In vitro and *in silico* methods to investigate and overcome drug resistance in cancer

Chiara Facciotto

Research Program in Systems Oncology (ONCOSYS)
Research Programs Unit
Biochemistry and Developmental Biology
Medicum
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 2, Haartmaninkatu 8, Helsinki,
on 13 August 2021, at 12 o'clock noon.

Helsinki 2021

Supervisor

Sampsa Hautaniemi, DTech, Professor
Research Program in Systems Oncology
Research Program Unit
Biochemistry and Developmental Biology
Medicum
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Rainer Lehtonen, PhD, Docent
Research Program in Applied Tumor Genomics
Research Program Unit
Department of Medical and Clinical Genetics
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Reviewers appointed by the Faculty

Kirsi Granberg, PhD, University Lecturer
Faculty of Medicine and Health Technology, Tampere University
Tampere, Finland

Tapio Pahikkala, PhD, Associate Professor
Department of Future Technologies, University of Turku
Turku, Finland

Opponent appointed by the Faculty

Marieke Kuijjer, PhD, Group Leader
Computational Biology and Systems Medicine group
Centre for Molecular Medicine Norway (NCMM), University of Oslo
Oslo, Norway

The Faulty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Cover illustration created with Biorender.com.

Series No. 41/2021
ISSN 2342-3161 (print)
ISSN 2342-317X (online)
ISBN 978-951-51-7426-0 (print)
ISBN 978-951-51-7427-7 (online)
`http://ethesis.helsinki.fi`
Unigrafia Oy
Helsinki 2021

To my family and friends, who supported me every step of the way.

Success is not final, failure is not fatal:
it is the courage to continue that counts.
Winston S. Churchill

Abstract

Overcoming drug resistance in cancer is one of the most pressing issues in oncology. The last century saw a dramatic increase in the discovery of new cancer therapies, so much so that chemotherapeutic agents and immunotherapies are now, alone or in combination, the backbone of treatment for many cancers. Despite the increased rate of treatment success brought by these regimens, cancer patients can become resistant to these drugs. This leads to disease relapse, hindering patient survival. Drug resistance remains the primary cause of death in most advanced-stage cancer patients.

The molecular mechanisms responsible for the development of a resistance phenotype in cancer cells are complex and include both genetic and epigenetic alterations. Since drug resistance is a multifactorial phenomenon, we used a systems biology approach to investigate it on different fronts. Specifically, we developed a high-throughput drug screening method to test new drug combinations, identifying epigenetic inhibitors able to sensitize lymphoma cells to doxorubicin. We also implemented a bioinformatic pipeline which combines multiple omics data to identify genes and pathways driving platinum response across multiple cancers. We then developed a method to compute differential methylation between cancer samples with varying and unknown tumor purity, which we used to investigate DNA methylation changes linked to drug resistance in ovarian cancer and lymphoma. Finally, we created a workflow management system to build complex bioinformatic pipelines and aid researchers in the analysis of high-throughput biomedical data.

By combining laboratory biology experiments and computational analyses, we gained a broader understanding of the cellular mechanisms behind immunochemotherapy failure. Moreover, we were able to identify novel biomarkers associated with platinum response in multiple cancers, as well as new drug combinations able to overcome immunochemotherapy resistance in lymphoma cells. The *in vitro* and *in silico* methods presented in this thesis can not only assist researchers in the cancer field, but are broadly applicable to other fields of biomedical research. Overall, this work is an important stepping stone in both understanding and overcoming drug resistance in cancer, and has great potential to improve outcomes for cancer patients in the future.

Sommario

Il superamento delle resistenze alle terapie nei malati di cancro è uno dei problemi più urgenti in oncologia. Durante lo scorso secolo, il numero delle terapie contro il cancro disponibili é aumentato considerevolmente, tanto che gli agenti chemioterapici e le immunoterapie sono ora, da soli o in combinazione, alla base del trattamento di molti tumori. Nonostante questi regimi terapeutici abbiano un maggiore tasso di successo rispetto all'uso di agenti singoli, i malati di cancro possono comunque sviluppare una resistenza a queste associazioni di farmaci. Questo porta alla recidiva della malattia, ostacolando la sopravvivenza del paziente. La resistenza ai farmaci rimane la principale causa di morte nella maggior parte dei pazienti oncologici in stadio avanzato.

Complessi meccanismi molecolari, che comprendono alterazioni sia genetiche che epigenetiche, sono responsabili dello sviluppo delle resistenze alle terapie nelle cellule tumorali. Poiché la resistenza ai farmaci è un fenomeno multifattoriale, abbiamo utilizzato un approccio di biologia dei sistemi per indagarlo su diversi fronti. Nello specifico, abbiamo sviluppato un metodo di high-throughput screening per testare nuove combinazioni di farmaci, identificando inibitori epigenetici in grado di sensibilizzare le cellule di linfoma alla doxorubicina. Abbiamo anche implementato una pipeline bioinformatica che combina dati omici per identificare geni e pathway molecolari che determinano la risposta al platino in diversi tipi di cancro. Abbiamo quindi sviluppato un metodo per stimare la differenza dello stato di metilazione del DNA tra campioni di cancro con purezza del tumore variabile e sconosciuta. Abbiamo poi usato questo metodo per studiare i cambiamenti della metilazione del DNA legati alla resistenza ai farmaci nel tumore alle ovaie e nel linfoma. Infine, abbiamo creato un workflow management system per costruire pipeline bioinformatiche complesse e aiutare i ricercatori nell'analisi di larghe quantitá di dati biomedici.

Combinando esperimenti biologici di laboratorio e analisi computazionali, abbiamo acquisito una più ampia comprensione dei meccanismi cellulari alla base del fallimento dell'immunochemioterapia. Inoltre, siamo stati in grado di identificare nuovi biomarcatori associati alla risposta al platino in diversi tipi di tumori, nonché nuove combinazioni di farmaci in grado di superare la resistenza all'immunochemioterapia nelle cellule di linfoma. I metodi *in vitro* e *in silico* presentati in questa tesi possono non solo aiutare i ricercatori nel campo oncologico, ma sono applicabili anche in altri campi della ricerca biomedica. Nel complesso, questo lavoro è un importante trampolino di lancio sia nella comprensione che nel superamento delle resistenze ai farmaci antitumorali, così da permettere ai pazienti oncologici un outcome e una qualità di vita migliori.

Contents

Abbreviations

ABC	Activated B-cell
aCGH	Array comparative genomic hybridization
Anduril	ANalysis of Data Using Rapid Integration of aLgorithms
ANXA9	Annexin A9
ARC	Activity Regulated Cytoskeleton Associated Protein
ARID1A	AT-Rich Interaction Domain 1A
Asc	HGSOC sample from ascites
ATP	Adenosine triphosphate
Bi-Seq	Bisulfite sequencing
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
BRCA1	Breast And Ovarian Cancer Susceptibility Protein 1
BRCA2	Breast And Ovarian Cancer Susceptibility Protein 2
BRD	Bromodomain
C19orf33	Chromosome 19 Open Reading Frame 33
CAR	Chimeric antigen receptor
Cs	Cytosine nucleotides
cDNA	Complementary DNA
CD20	Cluster of differentiation 20
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CGI	CpG island
CNA	Copy number alteration
COAD	Colon adenocarcinoma
CREBBP	Histone Lysine Acetyltransferase CREB Binding Protein
CpG	CG dinucleotide (i.e. a cytosine nucleotide followed by a guanine nucleotide in 5'$\rightarrow$ 3' direction)
DEG	Differentially expressed gene
DLBCL	Diffuse large B-cell lymphoma
DMG	Differential methylated gene
DMML	Differential methylation using a maximum-likelihood
DMP	Differential methylated promoter
DMS	Differential methylated CpG site
DNA	Deoxyribonucleic acid
DNAmeth	DNA methylation
DNMT	DNA methyltransferase
DSS	Dispersion shrinkage for sequencing data, a software to estimate differential methylation
DSS	Drug sensitivity score
ECI2	Enoyl-CoA Delta Isomerase 2
ENCODE	Encyclopedia of DNA Elements
EOC	Epithelial ovarian cancer
ESCA	Esophageal carcinoma
EZH2	Enhancer Of Zeste 2 Polycomb Repressive Complex 2 Subunit

FBXO17	F-Box Protein 17
FDA	US Federal Drug Administration
GBM	Glioblastoma multiforme
GCB	Germinal center B-cell
GDC	Genomic data commons
H3K27me3	Trimethylation of the 27th lysine residue of a histone H3 protein
H3K4me3	Trimethylation of the 4th lysine residue of a histone H3 protein
H3K27ac	Acetylation of the 27th lysine residue of a histone H3 protein
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HDACi	Histone deacetylase inhibitor
HDM	Histone demethylase
HGSOC	High grade serous ovarian cancer
HMT	Histone methyltrasferase
HNSC	Head and neck squamous cell carcinoma
HR	Homologous recombination
HRD	Homologous recombination deficiency
HTS	High-throughput screening
IC50	Half maximal inhibitory concentration of a compound or drug
LGALS3BP	Galectin 3 Binding Protein
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
mAb	Monoclonal antibody
Mes	HGSOC sample from intestine
MESO	Mesothelioma
MLF1	Myeloid Leukemia Factor 1
MOABS	Model-based Analysis of Bisulfite Sequencing data
mRNA	Messenger RNA
miRNA	Micro RNA
MRPS21	Mitochondrial Ribosomal Protein S21
MS4A1	Membrane-Spanning 4-Domains, Subfamily A, Member 1
NGS	Next generation sequencing technologies
NHEJ	Non-homologous end joining
NHL	Non-Hodgkin lymphoma
Ome	HGSOC sample from omentum
OS	Overall survival
OV	Ovarian serous cystadenocarcinoma
Ova	HGSOC sample from ovary
OvCa	Ovarian cancer
PAAD	Pancreatic adenocarcinoma
PCR	Polymerase chain reaction
Per	HGSOC sample from peritoneum
PFS	Progression free survival
QC	Quality control
R-CHOP	Rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone
READ	Rectum adenocarcinoma

RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RNAi	RNA interference
RRBS	Reduced representation bisulfite sequencing
scRNA-Seq	Single-cell RNA sequencing
SEC62	SEC62 Homolog, Preprotein Translocation Factor
siRNA	Small interfering RNA
SIRT1	Sirtuin 1
SNP	Single-nucleotide polymorphism
SOX17	SRY-Box Transcription Factor 17
STAD	Stomach adenocarcinoma
STAR	Spliced Transcripts Alignment to a Reference
Ts	Thymine nucleotides
TBS	Targeted bisulfite sequencing
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TGCT	Testicular germ cell tumors
TGS	Third generation sequencing technologies
TM4SF1	Transmembrane 4 L Six Family Member 1
TP53	Tumor Protein P53
TSPYL5	Testis-Specific Y-Encoded-Like Protein 5
Tub	HGSOC sample from Fallopian tubes
Tum	HGSOC sample from unknown location
Us	Uracil nucleotides
UCS	Uterine carcinosarcoma
UCEC	Uterine corpus endometrial carcinoma
WES	Whole exome sequencing
WGBS	Whole genome bisulfite sequencing
WHO	World Health Organization
WMS	Workflow management system

Publications

Publication I **Chiara Facciotto***, Julia Casado*, Laura Turunen, Suvi-Katri Leivonen, Manuela Tumiati, Ville Rantanen, Liisa Kauppi, Rainer Lehtonen, Sirpa Leppä, Krister Wennerberg, Sampsa Hautaniemi.
Drug screening approach combines epigenetic sensitization with immunochemotherapy in cancer.
Clinical Epigenetics **11**, 192 (2019)

Publication II **Chiara Facciotto**, Laura Lehtinen, Antti Häkkinen, Julia Casado, Ville Rantanen, Jaana Oikkonen, Veli-Matti Isoviita, Jani Saarela, Rainer Lehtonen, Sampsa Hautaniemi.
Identification of novel platinum-resistance biomarkers and therapeutic targets through integration of multi-omics data from 1,500 cancer patients.
Submitted

Publication III Antti Häkkinen, Amjad Alkodsi, **Chiara Facciotto**, Kaiyang Zhang, Katja Kaipio, Sirpa Leppä, Olli Carpén, Seija Grénman, Johanna Hynninen, Sakari Hietanen, Rainer Lehtonen, Sampsa Hautaniemi.
Identifying differentially methylated sites in samples with varying tumor purity.
Bioinformatics **34**, 18 (2018)

Publication IV Alejandra Cervera, Ville Rantanen, Kristian Ovaska, Marko Laakso, Javier Nuñez-Fontarnau, Amjad Alkodsi, Julia Casado, **Chiara Facciotto**, Antti Häkkinen, Riku Louhimo, Sirkku Karinen, Kaiyang Zhang, Kari Lavikka, Lauri Lyly, Maninder Pal Singh, Sampsa Hautaniemi.
Anduril 2: upgraded large-scale data integration framework.
Bioinformatics **35**, 19 (2019)

* equal contribution

In addition to the publications, unpublished results on DNA methylation in ovarian cancer are presented in the thesis.

Publications included in other theses

Publication I was included in the thesis of Julia Casado (Proteogenomics methods for translational cancer research, Helsinki 2021)

Author's contributions

Publication I Conceptualized the study, designed and optimized the drug screening protocol, implemented the drug screening experiments, built a pipeline to analyze the results, made the figures and wrote the manuscript (all together with JC). Processed RNA-Seq data, built the results explorer website.

Publication II Contributed to study conceptualization, designed the bioinformatic pipeline, integrated TCGA data, performed survival analysis, co-designed and analyzed the RNAi experiment, made the figures and wrote the manuscript.

Publication III Processed the ovarian cancer DNA methylation data used to test the method. Contributed to the manuscript.

Publication IV Contributed to the Anduril 2 project by designing, implementing and testing components, mainly related to the analysis of bisulfite sequencing and drug screening data. Contributed to the manuscript.

Thesis Contribution

Processed and analyzed the ovarian cancer DNA methylation data.

1 Introduction

Cancer is a group of over 100 different diseases, all characterized by the abnormal growth of cells which, due to mutations or other molecular alterations, have become neoplastic [1]. The World Health Organization (WHO) reported that worldwide, 18.1 million new cases of cancer were diagnosed and an estimated 9.6 million people died from cancer in 2018 [2]. Moreover, WHO also estimated that in 2010 the total annual economic cost of cancer at the global level was around 1.16 trillion US$ [3].

With the advent of modern medicine, the 19th and 20th centuries saw both an increase in cancer cases (due to improved overall health and an aging population) and also a boom in the development of new cancer therapies. However, the plastic nature of cancer allows the disease to develop resistance to treatments such as chemotherapies and immunotherapies. Hence, overcoming drug resistance in cancer to improve patient survival, reduce side effects, and decrease the cost of unsuccessful treatments is one of the major health challenges of the 21st century.

The molecular mechanisms at the root of drug resistance in cancer are complex and multivariate, often involving genetic and epigenetic aberrations [4]. Hence, a multidisciplinary approach is needed to better understand therapy failure, develop strategies to overcome it, and identify predictive biomarkers to avoid administering the wrong treatments to cancer patients. Thanks to modern high-throughput technologies, we are able to generate an unprecedented amount of molecular data, which represent a great opportunity but also present a major challenge. Correctly analyzing and interpreting such data to separate meaningful information from noise constitutes a major bottleneck in advancing our understanding of the molecular mechanisms behind cancer and other diseases [5]. Hence, tailored computational tools and pipelines combining multiple data layers are needed to advance our understanding of the field.

Diffuse large B-cell lymphoma (DLBCL) and high-grade serous ovarian cancer (HGSOC) are among the most aggressive forms of lymphomas and ovarian cancers, respectively. Even though standard treatments exist for these cancers, a significant proportion of patients still develop drug resistance [6, 7]. In this thesis, we used a systems biology approach to combine laboratory biology experiments and computational analyses to gain a more comprehensive understanding of the cellular mechanisms behind immunochemotherapy resistance. In Publication I, we identified epigenetic inhibitors able to reprogram DLBCL cells, sensitizing them to doxorubicin and rituximab. In Publication II, we studied the pan-cancer genetic and epigenetic landscape of patients treated with platinum-based chemotherapy

and, through a multi-omics integration analysis, we discovered novel biomarkers associated with platinum response. In Publication III, we developed a computational method to estimate differential methylation between cancer samples with varying tumor purities and tested it on DLBCL and HGSOC samples. In Publication IV, we present a bioinformatic workflow management system to implement easily reproducible computational pipelines, which was used to systematize and parallelize all data analyses carried out in Publications I and II.

This thesis proceeds with a literature review of the molecular mechanisms behind cancer and its resistance to standard therapies, with particular focus on DLBCL and HGSOC. Additionally, I present an overview of high throughput technologies used in Publications I-IV, including drug screenings, RNA sequencing and bisulfite sequencing, followed by a summary of the main methods and results developed to investigate drug resistance.

2 Cancer

In this section, I explore the molecular mechanisms responsible for the transformation of healthy cells into cancerous ones, as well as current cancer treatment options, their limitations and ways to overcome drug resistance in cancer.

2.1 Biology of cancer

Human bodies are made of hundreds of different types of cells, and anomalies in any of them can lead to different cancers. The biggest risk factor for cancer is aging. Over half of cancers manifest in people older than 70 [8], because the older a person is, the higher the chances are that their DNA has accumulated random mutations that can turn a healthy cell into a cancerous one. Other risk factors include exposure to UV light, radiation, dangerous chemicals, certain viruses (such as the human papillomavirus), an unhealthy diet, and smoking: all factors that can induce mutations in our genome [9].

Every person develops cancer cells throughout their life, but the immune system is generally very good at detecting and eliminating them. However, when cancer cells are able to escape immune detection and start to replicate, they can form malignant masses and spread across the body. In order to uncover the cascade of molecular changes that lead to cells becoming cancerous, it is important to first understand how healthy cells develop and maintain their own identity.

2.1.1 Regulation of gene expression and cellular identity

All cells from the same individual contain the same DNA sequence, the blueprint of the human body. Genes represent the portion of the information encoded in our genome that can be transcribed into RNA. A subset of genes, accounting for less than 2% of the human genome, are transcribed into messenger RNA (mRNA) and then further translated into proteins, the major building blocks of our cells. However, only a portion of genes are expressed in each cell: this is how the human body can generate different cell types that carry out very specialized functions and adapt to environmental stimuli.

A variety of intra- and extra-cellular signals work in concert to regulate which portion of the genome is transcribed, hence defining the cell's identity. Some of these mechanisms can activate or repress the transcription of a gene. These include the binding of transcription factors (TFs) to regulatory regions [10], as well as epigenetic mechanisms (such as DNA methylation and histone modifications)

which affect the chromatin structure [11, 12] and regulate which positions of the genome can be accessed by activating TFs, which in turn initiate transcription. Other mechanisms, such as RNA interference (RNAi) [13], can fine tune the expression level of a gene after it has been transcribed.

DNA methylation

Cytosines in the DNA can become methylated when a methyl group ($-CH_3$) binds to the C5 position of the nucleotide, turning it into a 5-methylcytosine. Although all cytosines have the potential to acquire a methyl group, in humans DNA methylation generally occurs at CpG sites, *i.e.* at cytosines that are followed by a guanine (Figure 1A). Specific enzymes, known as DNA methyltransferases (DNMTs), are responsible for adding the methyl group to the cystosine substrate [14]. Specifically, the DNMT3a and DNMT3b enzymes are responsible for *de novo* methylation, while DNMT1 maintains existing DNA methylation patterns during cell replication.

DNA methylation plays a different regulatory role depending on the genomic region where it is located [15]. Transposable elements, centromeres and other repetitive regions are usually methylated in order to ensure genome stability. There are also regions in the genome with a high concentration of CpG sites, which are known as CpG islands (CGIs) [16]. CGIs are generally unmethylated and approximately 60-70% of promoter regions in the human genome contain a CGI. Methylated promoters are usually associated with silencing of the related gene, while methylation in the body of a gene correlates with increased expression. DNA methylation can also mediate whether or not transcription factors or other proteins bind to the DNA [17, 18] and has been linked to regulation of RNA splicing [19].

Histone modification

Histones are proteins involved in the packaging of the DNA into chromatin. DNA strands wrap around histones, forming structures called nucleosomes [20]. There are five type of histones (H1, H2A, H2B, H3 and H4) and the N-tail of these proteins can acquire post-translational chemical modifications such as methylation, acetylation, phosphorylation, and ubiquitination (Figure 1B). These chemical groups are added and removed by specific enzymes [21]. For instance, histone methyltransferases (HMTs) are responsible for adding one or more methyl groups to histone tails, while histone demethylases (HDMs) remove such chemical groups. In the same way, histone acetyltransferases (HATs) and deacetylases (HDACs) are responsible for adding and removing acetyl groups. There are also enzymes that bind to histones when they recognize specific histone marks. For instance, chromodomain proteins bind to histones when a methyl mark is present, while bromodomain proteins (BRDs) recognize acetyl marks.

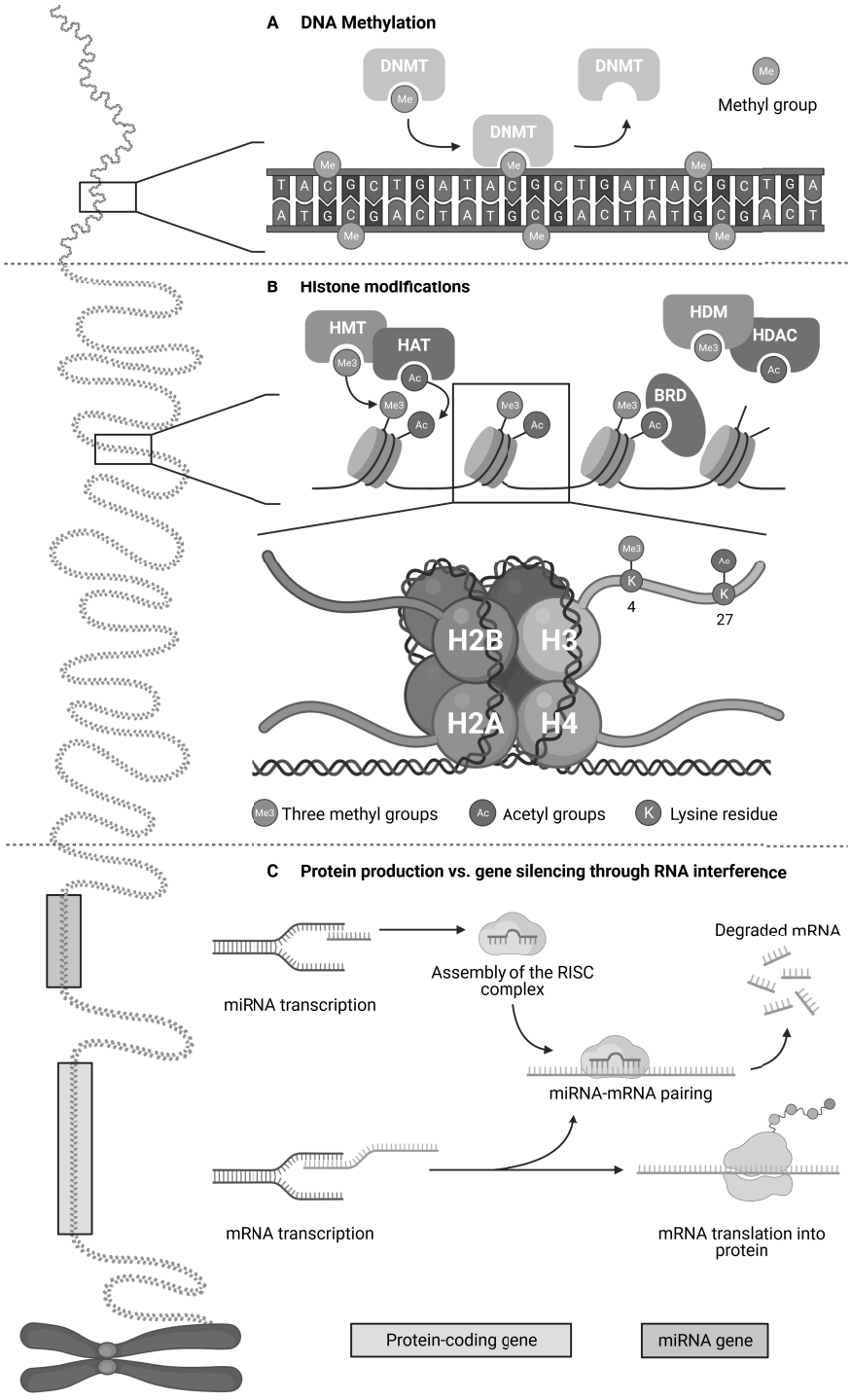

Figure 1: Cellular mechanisms of gene expression regulation. Caption continues in the next page. Image created with BioRender.com

Figure 1: (A) DNMT enzymes add methyl groups to the DNA at CpG sites. (B) Amino acids, such as lysines (Ks), forming the tails of histone proteins can be post-translationally modified via the addition of chemical groups. HMT and HAT enzymes add methyl and acetyl groups, respectively. Other enzymes like BRD can bind to histones, recognizing specific modifications. HDM and HDAC enzymes can remove methyl and acetyl groups from histone tails. (C) miRNAs are transcribed from DNA and, together with a set of proteins, form the RISC complex. miRNAs can then bind to mRNA molecules, which are then degraded by the RISC complex.

Histone modifications can alter the chromatin structure and play a key role in regulating gene expression. For instance, the trimethylation of the 27th lysine residue of a histone H3 protein (H3K27me3) is associated with gene silencing when it occurs in nucleosomes located in a gene promoter, while the trimethylation of the 4th lysine residue of a histone H3 protein (H3K4me3) in the same region correlates with actively transcribed genes [22]. Acetylation of the 27th lysine residue of a histone H3 protein (H3K27ac) alters the chromatin structure to make DNA more accessible to TFs and has also been associated with active enhancers [23].

RNA interference

RNAi is a mechanism of post-transcriptional gene regulation driven by two classes of small non-coding RNA species: microRNAs (miRNAs) or small interfering RNAs (siRNAs) [13]. MiRNAs and siRNAs are antisense RNAs, which means that they are complementary to specific mRNA molecules. These small RNA molecules can recruit specific enzymes to form the RNA-induced silencing complex (RISC) and guide it to the small RNA's complementary mRNA molecule, which is then cleaved and degraded, thus decreasing its expression level (Figure 1C). Even though mRNA degradation is the primary mechanism through which miRNAs affect gene expression, protein production can also be regulated via translational repression [24]. In this case, miRNAs interact with one of the complex responsible for mRNA maturation, preventing the mRNA molecule from reaching the conformation required for translation to take place.

Today, artificially designed siRNA molecules can be used to downregulate the expression of target genes *in vitro* via RNAi [25]. This allows us to observe how a cell's behavior can change with or without a certain protein.

2.1.2 When things go wrong - Cancer initiation and progression

When the mechanisms described above work correctly, healthy cells are in equilibrium with each other and the organism thrives. However, alterations in our DNA can disrupt this equilibrium, causing cells to acquire neoplastic characteristics. The

Figure 2: Dysregulation of gene expression though DNA methylation of cytosines in the promoter region or alterations in the number of copies of a gene. An increase in gene expression is generally observed when the promoter of the gene is unmethylated and/or the portion of the DNA encoding the gene has been amplified (*i.e.*, more than two copies are present). Conversely, a methylated promoter and/or the deletion of one or both copies of the gene would lead to a decreased expression of the gene.

term neoplasm was first introduced in the 19th century and comes from the Greek words *neos* or "new" and *plasma* or "formation". Today "neoplasm" indicates a group of cells displaying abnormal and excessive growth, uncoordinated with the surrounding tissues. The immune system is generally able to detect neoplastic cells, but if such cells escape immune detection they continue to replicate and eventually form a mass, called a tumor. If the growth remains localized, the tumor is considered benign. However, if the neoplastic cells start to spread to other parts of the body, damaging other tissues, the tumor is then classified as malignant. Malignant tumors, also known as cancers, can be fatal for the individual, especially if left untreated. Every cell type in our body has the potential to become cancerous, and cancers originating from different tissues may undergo a different progression and require different treatment strategies. Despite these differences, cancer cells share a set of mechanisms, known as the hallmarks of cancer [26], which allow them to replicate, proliferate, and spread around the body.

The majority of cancers (90-95%) are caused by somatic mutations, but a small portion of cancers (5-10%) can be inherited due to germline mutations, *i.e.* mutations in a sperm cell or an egg cell, which can be passed from one generation to the next [27]. The longer cancer cells survive, the more their genomes acquire alterations, such as point mutations and/or amplifications and deletions of entire chromosomal sections. This can lead to a chain reaction that affects the mechanisms regulating gene expression, further advancing tumor progression. Alterations in the number of copies of a gene, as well as the disruption of epigenetic patterns, are among the main aberrations that can lead to transcriptional changes, as shown in Figure 2.

Gene copy number alterations

Healthy cells posses two copies of each gene, one inherited from each parent. However, cancer cells are characterized by genomic instability [28], which is the accumulation of structural aberrations in the genome. These aberrations can be as small as the mutation of a single nucleotide [29, 30], but can also affect broader portions of the DNA via amplifications and deletions of gene-containing chromosomal regions, or even entire chromosomes. Alterations in the number of copies of a gene affect its expression [28], leading to changes that can promote tumor progression. Moreover, copy number alterations (CNAs) of certain genes have also been associated with resistance to certain cancer drugs. For instance, *ERBB2* [31] and *CCNE1* [32] amplifications have been linked to platinum resistance.

Dysregulation of the epigenome

Healthy cells can flexibly adapt to different stimuli by altering their epigenomes [33]. However, dysregulation of epigenetic patterns is a common feature of cancer cells, which can alter cells' proliferation rate as well as gene expression. Epigenetic changes can even decrease the differentiation state of cancer cells, allowing them to acquire stemness properties that in turn make them more plastic, increasing their ability to develop drug resistance mechanisms.

Disruption of DNA methylation patterns is known to be a major feature of cancer cells, manifesting as a global hypomethylation of the genome, but also as localized hypermethylation of CpG islands and promoter regions [34]. Global DNA hypomethylation leads to genomic instability, while hypermethylation of gene promoters might silence tumor suppressor genes. For instance, the silencing of the *BRCA1* gene through promoter methylation is a known mechanism associated to breast and ovarian cancer [35].

Patterns of histone modifications are also altered in cancer, leading to abnormal chromatin conformation and to the activation or inactivation of certain genes [36]. Examples of epigenetic dysregulation affecting histone modifications include mutations of the HMT gene *EZH2* in some forms of lymphoma [37], which leads to altered patterns of H3K27me3, and overexpression of the HDAC *SIRT1* in prostate cancer [38] which affects histone acetylation.

Nucleosome repositioning [39] and altered networks of non-coding RNAs [40] are other epigenetic mechanisms commonly disrupted in cancer.

2.1.3 Cancer heterogeneity

Due to the chaotic nature of tumor progression, a tumor mass is a very heterogeneous collection of cells (Figure 3). This heterogeneity appears on two levels:

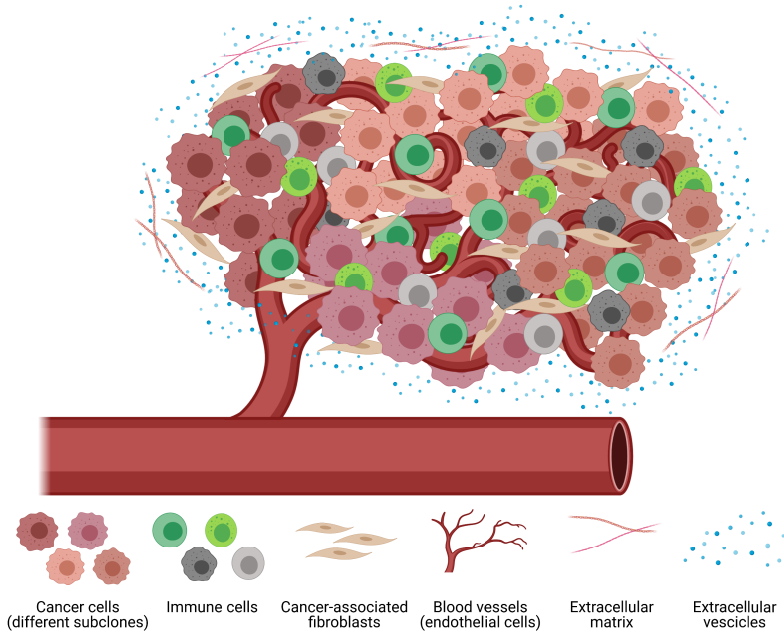

Figure 3: A tumor mass is a heterogeneous collection of cells embedded in extracellular matrix. Since cancer cells proliferate quickly and accumulate structural aberrations, different subclones of cancer cells are present in the same tumor. At the same time, the tumor tissue is infiltrated by other cell types, including endothelial cells forming blood vessels, immune cells, and cancer-associated fibroblasts. Cells communicate by releasing extracellular vescicles and other cell signaling molecules. Image created with BioRender.com

clonal heterogeneity and tumor microenvironment composition.

Clonal heterogeneity occurs when cell replication leads to the accumulation of genetic and epigenetic changes that create diverging groups of cancer cells, known as subclones [41, 42]. Factors driving this type of heterogeneity are, for instance, hypoxia and inflammation.

The other level of heterogeneity is derived from the different cell types that are part of the tumor microenvironment [43]. Cancer cells can hijack the components of these other cell types to promote their own growth and survival. For instance, a tumor requires blood vessels that can deliver nutrients to its cells, so endothelial cells are recruited during neoangiogenesis (*i.e.* the new growth of blood vessels) [44]. Moreover, immune cells are also able to infiltrate the tumor tissue and, by releasing specific chemicals, can support cancer growth [45]. Cancer-associated fibroblasts are another important part of the tumor microenvironment [46], where they provide a structural framework for the neoplasm and synthesize extracellular matrix, which in turn supports the transport of chemical signals, extracellular

vesicles and nutrients across different cells.

2.2 Diffuse large B-cell lymphoma

Lymphomas are the most common type of blood cancer, specifically affecting a type of white blood cell known as a lymphocyte. Lymphocytes are the main component of our immune system. They travel through the blood and the lymphatic system to defend the body by identifying and neutralizing pathogens, like bacteria and viruses, clearing the organism of toxic and allergenic substances, and eliminating neoplastic cells [47]. In particular, B cells are a specific subset of lymphocytes responsible for the production of antibodies, the proteins that are able to detect both microbes and unhealthy cells, and hence are fundamental to fighting infections. However, like every other human cell, B cells can also become cancerous, spreading through the body and growing in the lymph nodes, bone marrow, spleen, and other organs [48].

Lymphomas are generally classified as either Hodgkin lymphoma or non-Hodgkin lymphoma (NHL). The latter group includes about 90% of all lymphomas [49]. Worldwide, in 2018 alone 509,590 new NHL cases were diagnosed and 248,724 patients died due to NHL [50]. Diffuse large B-cell lymphomas (DLBCLs) are the most common type of NHL, accounting for 30-40% of NHL cases [51]. DLBCLs can be classified into two major subtypes based on the development stage of the B cells from which they originate: the germinal centre B-cell like (GCB) subtype, usually with better prognosis, and the activated B-cell like (ABC) subtype [51]. Most DLBCLs belong to one of these two subtypes, but about 10-15% of cases are unclassifiable.

Even though it can manifest at early age, the occurrence of DLBCL increases with age and the median age at diagnosis is 70 [52]. The incidence is also slightly higher in men [52].

Perturbations of the epigenetic landscape are strongly linked to DLBCL [53]. One example is the dysregulation of histone methylation patterns in DLBCL due to mutations of the *EZH2* gene, common in the GCB subtype [37].

2.3 High-grade serous ovarian cancer

Ovarian cancer is one of the most common and the most deadly type of gynecologic cancer, encompassing a collection of neoplasms originating from the ovaries or the Fallopian tubes. In 2018, 295,414 new ovarian cancers were diagnosed, and 184,799 women died of this disease worldwide[50]. About 90% of ovarian cancers have an epithelial origin, while the remaining 10% originate either from germ or

stromal cells and are less invasive. A small portion (~3%) of epithelial ovarian cancers (EOC) are mucinous, while about 20% of all EOCs are associated with endometriosis and are classified as either clear cell or endometrioid carcinomas [54]. The remaining EOC cases originate from epithelial cells of either the Fallopian tubes or the surface of the ovaries [55], and are divided into high-grade serous ovarian cancers (HGSOCs) and low-grade serous ovarian cancers.

HGSOC is the most common and aggressive form of EOC (~70% of cases). One of the causes of HGSOC is hereditary breast-ovarian cancer syndrome, an autosomal dominant genetic disorder caused by specific genetic mutations in the *BRCA1* and *BRCA2* genes. *BRCA1* and *BRCA2* are tumor suppressor genes involved in DNA repair pathways, including homologous recombination. Mutations in the *TP53* gene are also present in almost all HGSOC [56]. Chromosomal instability, homologous recombination deficiency (HRD), aneuploidy and somatic mutations are other typical features of HGSOC.

HGSOCs originate in the ovaries or Fallopian tubes (stage I) and spread first to other pelvic organs such as the uterus (stage II), and then gradually to organs or tissues within the peritoneal cavity (stage III) or even beyond, reaching the lungs and extra-abdominal lymph nodes (stage IV) [57]. The high mortality rate of HGSOC is due to the unspecific nature of the symptoms at the early stages of the disease, as well as the lack of accurate screening methods [58]. In fact, over 85% of HGSOC diagnoses occur at stage III or IV [59] and, by this time, the tumor is already highly heterogeneous, which increases the probability of developing a treatment-resistant subclone within the tumor mass.

2.4 Cancer therapies

Several treatment strategies are currently available to treat cancer, including surgery, chemotherapy, and radiotherapy, which are the most common and established methods. More recent therapies include immunotherapy and targeted therapy. These treatments can be used alone or in combination, depending on the type of cancer, its stage and location, and the particular medical history of each patient.

In this thesis, I focus mainly on cancers where chemotherapy is used in combination with surgery or with immunotherapy. Hence, more information on these three treatment options can be found below.

Surgery. Surgery is the oldest treatment option for solid cancers [60] and documents reporting the surgical removal of tumors date back to ancient Egypt. However, it is only since the middle of the 19th century, when anaesthesia and antisepsis were introduced, that cancer surgery has become a widespread option to treat solid

malignancies. Not all tumors are operable, but when they are the surgical removal of the mass before cancer cells spread to other organs leads to a better prognosis for the patient. In the beginning of the 20th century, surgeons started performing increasingly radical surgeries, removing large amounts of tissue, but they realized that when the cancer has already spread to other organs, surgery alone does not improve patient survival. Since the emergence of other treatment options, surgery has mostly been used in combination with other therapies.

Chemotherapy. Chemotherapy was first introduced around the middle of the 20th century [61], and since then a plethora of chemotherapeutic agents have been approved for the treatment of different malignancies. Chemotherapy drugs are chemical compounds that target fast-growing cells by exploiting certain mechanisms, such as halting the cell replication cycle or damaging DNA. Since cancer cells are often unable to repair damaged DNA, this triggers a chain reaction leading to apoptosis.

Multiple chemotherapy agents targeting different molecular mechanisms can be combined into standard chemotherapy regimens, which specify the drugs to use, their dosage, and the frequency at which they should be administered to treat a specific cancer.

Compared to surgery, which is localized to a certain part of the body, chemotherapy is generally administered intravenously or orally, which means that the drug is transported throughout the whole body via our circulatory system. This makes it a more effective option for non-solid cancers or when the tumor has already metastasized. However, since chemotherapy does not target cancer cells specifically, this type of treatment is also damaging for healthy tissues and it generally causes more side effects than other treatment options.

Immunotherapy. Cancer immunotherapy is a branch of oncology that develops a broad range of therapeutic strategies that, rather than aiming to directly kill cancer cells, aims to make them recognizable to the immune system of the patient, which in turn will fight the cancer. Immunotherapy exploits the fact that neoplastic cells present specific antigens on their cell membranes.

A common and widespread strategy for passive immunotherapy employs artificial monoclonal antibodies (mAb), designed to recognize tumor antigens and recruit immune cells by mimicking natural antibodies [62]. Other types of passive immunotherapies include checkpoint inhibitors [63] and cytokine therapy [64]. Active immunotherapy options, like cancer vaccines [65] and the use of CAR T-cells [66], have also been developed in recent years.

2.4.1 Treating DLBCL

The standard of care for DLBCL patients consists of the immunochemotherapy regimen R-CHOP, which combines the monoclonal antibody rituximab with four chemotherapeutic agents (cyclophosphamide, doxorubicin, vincristine, and prednisone) [67]. Though most patients respond well to R-CHOP, about 30-40% relapse or continue to have refractory disease, and resistance to R-CHOP is still the most common cause of mortality among DLBCL patients [6].

Rituximab. Rituximab is a chimeric mouse/human mAb targeting the CD20 protein. CD20, encoded in the *MS4A1* gene, is expressed on the surface of B cells, making it an ideal therapeutic target for malignancies originating from this cell type. Rituximab was approved in 1997, making it the first therapeutic mAb approved for cancer treatment [68]. The addition of rituximab to the CHOP regimen led to a 10% increase in the 5-year overall survival of DLBCL patients [69, 70, 71, 72, 73]. Rituximab can recruit T cells [74], which in turn are able to kill cancer cells, or can directly induce apoptosis via the p38 MAP-kinase signaling pathway [75].

Doxorubicin. Doxorubicin belongs to a class of chemotherapeutic drugs known as anthracyclines [76], which are extracted from the *Streptomyces* bacterium. Doxorubicin is used to treat a wide range of cancers, including DLBCL, and it is generally used in combination with other chemotherapeutic agents. Doxorubicin intercalates into DNA molecules, where it inhibits the progression of topoisomerase II, an enzyme which relaxes supercoils in DNA for transcription and replication [77]. In this way it can halt DNA replication, consequently stopping the proliferation of cells (including the abnormal growth of cancer cells).

2.4.2 Treating HGSOC

The standard treatment for HGSOC patients consists of a cytoreductive surgery or "debulking" [78]. The surgeon proceeds to remove the ovaries, fallopian tubes, and uterus to resect all macroscopic tumor masses. This procedure is more successful in early-stage patients because the cancer has not yet spread to more distal parts of the body. To make sure to completely eradicate the disease, the majority of patients receive adjuvant chemotherapy after surgery. For the last 20 years, a combination of platinum-based drugs and taxanes has been the standard chemotherapy administered to HGSOC patients [78]. In recent years, the administration of neoadjuvant chemotherapy prior to debulking has also been investigated in the clinic, as an alternative for patients too sick to undergo surgery or whose tumor is too widespread to be operable [79]. In this case, the patient would receive neoadjuvant chemotherapy followed by interval debulking and additional

adjuvant chemotherapy.

Even though the platinum/taxane combination is currently the most effective chemotherapy regimen for HGSOC patients, $\geq$80% of them will eventually relapse at some stage [57].

Platinum-based drugs. Currently there are three platinum-based drugs approved worldwide [80]. Cisplatin was the first platinum compound approved as an anti-cancer drug in 1978, followed by the less toxic analog carboplatin in 1986 and oxaliplatin in 2004. Three other platinum compounds have been approved for clinical use in specific countries (*i.e.*, nedaplatin in Japan, lobaplatin in China, and heptaplatin in Korea), while many more are currently being developed or are already undergoing clinical trials.

Platinum-based drugs are absorbed by cells and transported to the nucleus, where they cause DNA damage by crosslinking with purine bases and inducing apoptosis [81].

In addition to being the main treatment for HGSOCs, platinum drugs are widely used in combination therapies for several other cancers. In fact, platinum is administered to half of patients receiving chemotherapy [82], and to 10-20% of all cancer patients according to the National Cancer Institute [83].

2.5 Drug resistance in cancer

Despite our increased understanding of the molecular mechanisms behind cancer and the plethora of treatments developed in the last century, resistance to such therapies still hinders patient survival.

Drug resistance is defined as intrinsic when the heterogeneous tumor mass contains subclones that are resistant to a certain therapy before the therapy itself is administered (Figure 4A). This type of resistance is more common in cancers diagnosed at a later stage as the tumors have had more time to accumulate genetic and epigenetic changes, making them more likely to develop a resistant phenotype. Drug sensitivity can also be lost during treatment, when the therapy administered to the patient drives the formation of resistant subclones (Figure 4B). For instance, DNA-damaging chemotherapy agents might cause mutations that confer resistance to that same drug.

The mechanisms behind drug resistance in cancer are complex and still not fully understood. However, genes responsible for the absorption, distribution, metabolism, and excretion of drugs have been implicated. Alterations of these genes can lead to reduced drug intake, increased drug efflux, or even inactivation of the drug itself, all mechanisms that can lead to a decrease in drug efficacy (Figure 4C). Moreover,

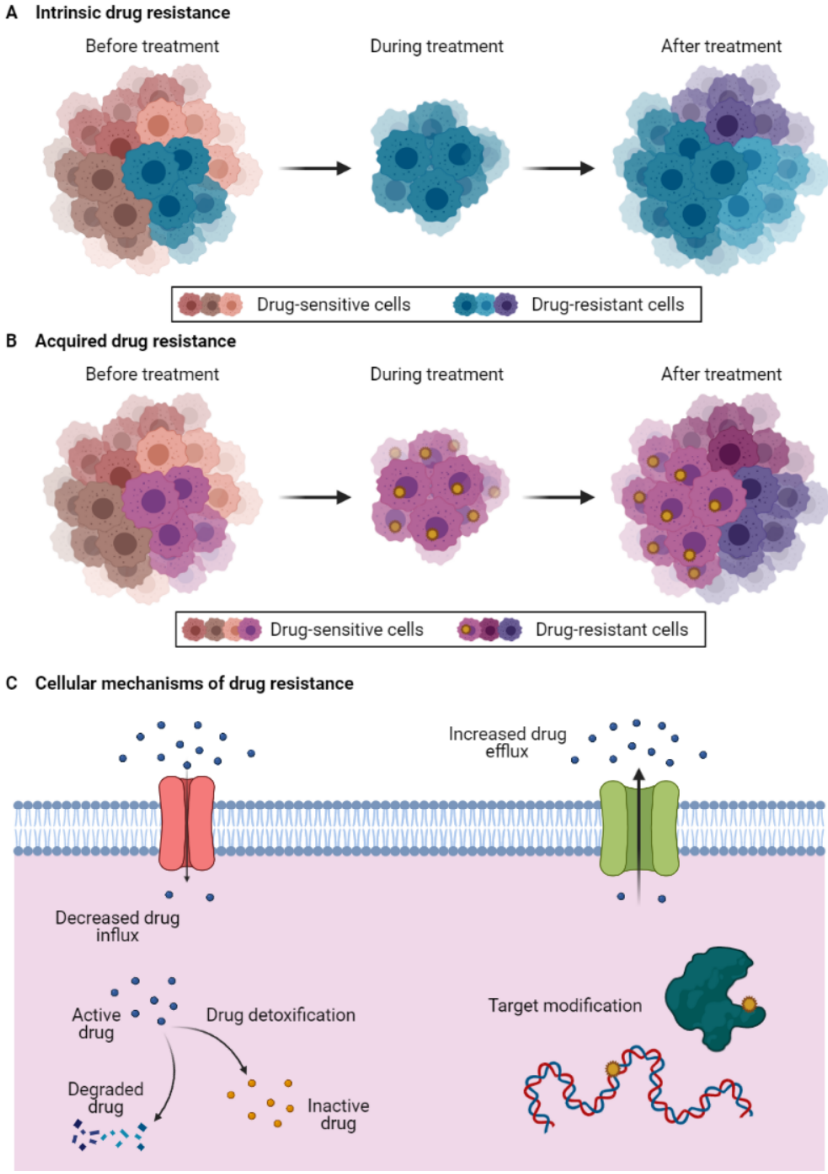

Figure 4: Mechanisms of drug resistance. (A) Intrinsic drug resistance occurs when one or more subpopulations of cancer cells already developed a resistant phenotype before the treatment is administered. (B) Acquired drug resistance occurs when the treatment causes changes in one or more cancer cells, which makes them insensitive to the effect of the drug. (C) Cellular mechanisms of drug resistance such as decreased drug influx, increased drug efflux, and drug detoxification via inactivation or degradation contribute to decreasing the drug concentration inside the cell, hence reducing its cytotoxic effect. Resistance may also arise when the structure of the drug target is altered. Image created with BioRender.com

since drugs reach cancer cells via the circulatory system, cells located in hypoxic regions of the tumor mass might survive treatment simply because the therapy is not able to reach them [84].

Changes in the structure, abundance or location of a drug target can also affect the response to a specific therapy (Figure 4C). For instance, downregulation of the *MS4A1* gene (due to genetic or epigenetic factors) or defects in the transport of CD20 to the cell membrane have been linked to rituximab resistance [85].

Certain cellular mechanisms might also nullify the effect of a drug. For example, cancer cells with functional DNA repair pathways (such as homologous recombination) are more likely to develop resistance to DNA-damaging agents like platinum-based drugs [86] and doxorubicin [87].

2.6 Predicting and overcoming drug resistance in cancer

Transitioning from single-agent therapies to combination therapies was one of the first approaches to overcoming drug resistance in cancer. However, even multi-drug regimens fail to completely eradicate cancer cells in an individual. Hence, new strategies to identify patients less likely to benefit from a certain therapy, and to overcome drug resistance are urgently needed.

2.6.1 Identification of biomarkers

Biomarkers are measurable indicators used to accurately and reproducibly classify a certain biological condition. To overcome drug resistance in cancer, better biomarkers are needed to both (i) diagnose tumors at earlier stages (when they are less likely to have developed drug resistance), and (ii) predict whether a patient will respond to a certain treatment, so therapies are administered only when the probability of success is high enough to justify exposing the patient to its side effects [88]. Different biological samples (*e.g.* blood, urine, soft tissues, tumor samples, etc.) and different data types (*e.g.* genetic [89], epigenetic [89], transcriptomic [90], proteomic [91], and imaging [92] data) can be used to identify biomarkers. For instance, mutations in the DNA repair genes *BRCA1* and *BRCA2* are biomarkers that can predict the efficacy of PARP inhibitors when treating ovarian cancer [93].

2.6.2 Restoring drug sensitivity via epigenetic reprogramming

In the last few decades, the plasticity of the epigenome has emerged as an appealing avenue to overcome drug resistance in cancer. Currently, a plethora of epigenetic inhibitors have been developed in order to block the activity of the main classes of

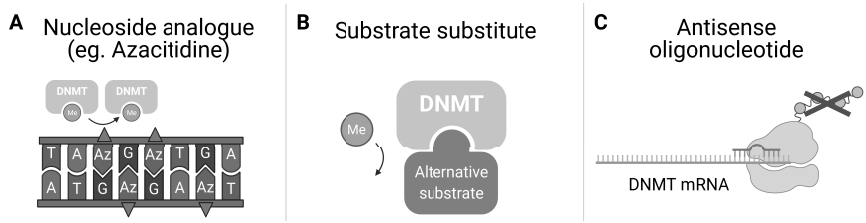

Figure 5: Different strategies to inhibit DNMT enzymes. (A) Cytidine analogues like Azacitidine are incorporated in the DNA during cellular replication and, because of their molecular structure, cannot be methylated, consequently inhibiting the action of DNMT enzymes. (B) Substrate substitutes targeting the active site of DNMT enzymes inhibit DNMTs from binding with methyl groups. (C) Antisense oligonicleotides targeting mRNA molecules encoding DNMTs can halt the production of these enzymes. Image created with BioRender.com

epigenetic enzymes, including DNMTs, HDACs, HATs, BRDs, HMTs, and HDMs. So far, only six epigenetic compounds have reached FDA approval [94]: the DNMT inhibitors azacitidine (in 2004) and decitabine (in 2006), and the HDAC inhibitors romidepsin (in 2004), vorinostat (in 2006), belinostat (in 2014), and panobinostat (in 2015). Other drugs, like valporic acid, have also been approved but not for their epigenetic effect.

Different strategies have been used to design compounds that are able to inhibit epigenetic enzymes. For instance, when aiming to inhibit DNMTs, one option is to design nucleoside analogues (*e.g.* azacitidine and decitabine) containing a modified cytosine ring that cannot be methylated (Figure 5A) [95]. During cell replication, these molecules are incorporated into the DNA strand, making it impossible for DNMTs to methylate the DNA containing the modified nucleosides. Another strategy used to inhibit DNMT enzymes consists of designing small molecule inhibitors (*e.g.* RG108 [96]) that are able to bind directly to the DNMT's catalytic region, hence preventing the binding between the DNMT enzyme and a methyl group (Figure 5B). Decreasing the abundance of a DNMT enzyme is also a way to inhibit its activity, and can for instance be achieved using antisense oligonucleotides (*e.g.* MG98 [97]) that bind to the enzyme's mRNA and prevent it from being translated into protein (Figure 5C).

Epigenetic inhibitors are an emerging treatment option for several diseases, including diabetes, cardiovascular and immune diseases, mental illnesses, anxiety disorders, and of course cancer [98]. The addition of epigenetic inhibitors to cancer regimens is being investigated, as in the case of the HMT inhibitor tazemetostat tested in combination with R-CHOP to treat newly diagnosed DLBCLs with poor prognosis [99]. However, recently it has been suggested that using epigenetic inhibitors as pretreatments might be a useful strategy to sensitize cancers that are

resistant to standard treatment. Successful chemosensitization through epigenetic reprogramming has already been shown *in vitro* in blood cancers such as acute lymphoblastic leukemia [100], but further investigation into this therapeutic option is needed before testing it in the clinic.

3 High-throughput biology

In the last few decades, several new high-throughput sequencing technologies have reshaped the landscape of biomedical research, shifting the scale from single genes to entire genomes. Consequently, novel computational tools tailored to process platform-specific data have also been developed. The widespread use of these technologies has drastically improved our understanding of the molecular mechanisms regulating healthy cells and driving tumor progression. In turn, this has led to the identification of several novel targets for cancer treatment, boosting the drug discovery pipeline.

The number of cancer drugs available on the market has been growing exponentially, and so has the number of compounds under development. Moreover, repurposing of existing drugs for cancer treatment requires assays that can screen thousand of compounds to select the most promising ones. Hence, high-throughput drug screening protocols were developed to efficiently test drug response in multiple cancer samples. By integrating information from sequencing and drug response data, we are now able to investigate therapy resistance in different cancers with an unprecedented level of accuracy.

3.1 High-throughput screenings

High-throughput screenings (HTSs) have become a standard assay which allows for the testing of multiple drugs and drug combinations using *in vitro* models, such as cell lines or organoids, or *ex vivo* patient samples [101]. Cells are seeded in multi-well microtiter plates and the use of liquid-handling robotics allows for the investigation of tens, hundreds, or even thousands of compounds at varying concentrations and in a short amount of time. A standard HTS protocol includes seeding the cells on the plate, adding the compounds to be tested in each well and, after a certain amount of time, measuring the proportion of cells that were killed by each compound.

When designing an HTS experiment, it is fundamental to select the right samples, and to design a comprehensive compound library. Plate effect errors, caused for instance by increased medium evaporation in the border wells of the plate, can also be accounted for by randomizing the plate design and by including replicates for each measurement. Controls are also extremely important when planning a drug screening experiment, because they allow researchers to normalize and interpret the effect of the compounds tested. A negative control is a substance that should not affect cells' viability. Dimethyl sulfoxide (DMSO) and water, commonly used to

dissolve drugs, are examples of suitable negative controls. A substance that would instead achieve maximal cytotoxic effect is used as positive control. An example of such a compound is benzethonium chloride, a commonly used antiseptic.

Different drug response quantification techniques are currently available to estimate the viability of cells after treatment. A common assay employs the CellTiter-Glo® reagent (Promega), which lyses the cells and generates a luminescent signal proportional to the amount of ATP, an indicator of metabolically active cells. The luminescence readout is then measured using a microplate reader.

Including multiple concentrations of the same compound in the screening generates dose response data, which are used to estimate the half maximal inhibitory concentration (IC50) of a compound, *i.e.* the dose of the compound needed to kill half of the cells present in the sample. The IC50 is used as the standard measure of the potency of a drug.

High-throughput screening can also be used to test how the knockdown of a certain gene affects cells' viability, their phenotype and/or their response to a specific therapy. These screenings exploit RNAi mechanisms by transfecting cells with artificial siRNA [25].

3.1.1 Analyzing dose response data

The standard computational analysis of dose response data obtained from HTS follows the steps summarized in Figure 6. First, HTS luminescence measurements generated by a microplate reader are annotated to mark what is measured in each well. The reliability of the readout is then estimated via standard quality control (QC) statistics, which include coefficient of variation for both positive and negative controls, Z' score, and strictly standardized mean difference [102]. If no severe bias or plate effect is detected, the luminescence readout is normalized using the negative controls (0% inhibition) and the positive controls (100% inhibition).

Normalized dose response data for a drug are fitted to a sigmoidal function. A common model used to fit the curve is the 4-parameter logistic function:

$$y = a + \frac{d-a}{1+(x/c)^{-b}} \tag{1}$$

where a and d represent the theoretical response when no drug is administered ($x = 0$) and when administering an infinite amount of drug ($x = +\infty$), respectively. c represents the IC50, and b is the slope factor of the curve at $x =$IC50. The R package drc [103] performs curve fitting and estimates the IC50 of a drug. Fitted curves are then used to estimate the effect of the drug on the cells, for instance by

Figure 6: Standard bioinformatic pipeline describing the analysis of dose response data obtained from a drug screening. Cell viability is estimated using a microplate reader which measures the luminescent signal emitted by the fraction of living cells in each well. The experimental conditions tested in each well are used to annotate the data. Next, QC statistics are estimated to assess the reliability of the experiment, and data are normalized using positive and negative controls. Data are then fitted to a sigmoidal function, which is used to estimate the effect of the drug.

computing drug sensitivity scores (DSSs) [104]. DSSs can then be used to compare the effect of different compounds and conditions on the same sample.

Software such as the online tool Breeze can analyze dose response data from QC to DSS calculations [105].

3.2 Next-generation sequencing technologies

The first sequencing technologies were developed by Sanger and his colleagues at Cambridge University in 1977 [106]. These technologies read DNA and RNA sequences at a single-base resolution, but were very low-throughput and quite costly. Consequently, since the mid-1990s hybridization-based microarrays became the standard high-throughput assays to quantify gene expression and measure genomic and epigenomic changes [107, 108]. However, despite being very powerful and efficient, microarrays are limited in the sense that they do not allow for measurement of *de novo* sequences, restricting investigation to those genes and DNA regions that are already known. In 2005, a new generation of sequencing technologies, also known as next-generation sequencing (NGS), was developed [109]. These technologies combined the precision of Sanger sequencing with the ability to produce a high-throughput readout. Although quite expensive at first, the costs of NGS assays has steadily decreased in the last decade, drastically changing the way work in fields like genomics, epigenomics, and transciptomics is conducted. By allowing researchers to investigate these molecular mechanisms at a single-base resolution and with unprecedented precision, NGS has undoubtedly increased our understanding of cellular biology.

Among the most common applications of NGS are whole genome and whole exome sequencing for the study of mutations and copy number alterations, RNA sequencing [110] to investigate changes in the transcriptome, bisulfite sequencing to map methylated cytosines across the genome, chromatin immunoprecipitation

sequencing (ChIP-Seq) to identify transcription factor binding sites and locate histone modifications, and an assay for transposase-accessible chromatin using sequencing (ATAC-Seq).

In the last ten years, a third generation of sequencing technologies (TGS), also known as long-read sequencing, was introduced [111]. Nanopore technologies [112] and other TGSs can sequence substantially longer reads than NGSs, producing genome assemblies of unprecedented quality. Moreover, TGSs can detect DNA methylation without the need to perform bisulfite conversion.

3.2.1 RNA sequencing

RNA sequencing (RNA-Seq) exploits NGS technologies to quantify the abundance of RNA transcripts and to estimate differential gene expression between samples. In addition, it also allows for the investigation of other mechanisms such as alternative splicing, gene fusion, and mutations.

The RNA-Seq protocol starts with the extraction of RNA molecules from the cells. Investigation can be done on the whole transcriptome (total RNA) or focus only on protein coding genes by selecting only mature mRNAs (by capturing RNA molecules that have a poly-A tail). In both cases, ribosomal RNAs, which account for over 80% of the RNA molecules inside a cell, are discarded. Size-selection is then carried out to identify specific RNA molecules, such as small non-coding RNAs. Next, RNA is fragmented and converted into cDNA to prepare the sequencing library. Adaptors are added at both ends of each cDNA molecule. Finally, the library is amplified via polymerase chain reaction (PCR) and sequenced.

In addition to the standard RNA-Seq assay, which measures the average transcript abundance from all cells in a bulk sample, a single-cell (sc) RNA-Seq protocol has also been developed [113]. ScRNA-Seq processes each cell individually, allowing researchers to obtain gene expression signatures from different cell types.

3.2.2 Bisulfite sequencing

Bisulfite sequencing (Bi-Seq) is a technology that allows for the genome-wide investigation of DNA methylation at a single-base level. Bisulfite conversion uses sodium bisulfite to deaminate unmethylated cytosines (Cs) into uracils (Us), which are then converted into thymines (Ts) during the PCR step. Methylated Cs are not affected by the bisulfite conversion reaction and can hence be detected once the reads are aligned to a reference genome. A T in the data mapping to a T in the reference genome is a T, a T mapping to a C is an unmethylated C, and a C mapping to a C is a methylated C.

Currently there are three protocols to perform Bi-Seq of bulk samples. Whole genome bisulfite sequencing (WGBS) [114] was the first method to be developed. Even though it allows for the analysis of all 29 million CpG sites found in the human genome, it was initially quite costly and required a high amount of starting genetic material, especially when aiming to produce data with high sequencing depth. To address this limitation, the reduced representation bisulfite sequencing (RRBS) [115] protocol was introduced. RRBS only sequences about 1-10% of the genome, targeting CpG-rich regions which include most CpG islands, promoters, gene bodies, other regulatory elements, and repetitive sequences. A third way to perform Bi-Seq, known as targeted bisulfite sequencing (TBS), is an enrichment-based procedure able to capture the methylation of CpGs located at specific genomic regions via hybridization through a probe-based system [116]. In this way, the experimental costs and the data size are both reduced with respect to WGBS, while ensuring consistent sequencing of all significant portions of the genome. A standard array targeting about 3.7 million CpGs is available, covering most promoters and other relevant genomic regions, but users can also customize the probes to target specific portion of the genome.

WGBS, RRBS, and TBS follow a very similar protocol. Briefly, DNA is extracted from the cells, purified from the proteins bound to it, and fragmented. WGBS and TBS use sonication to randomly fragment the DNA molecule, while RRBS cuts the genome at specific sites by using a restriction enzyme. The fragments undergo end repair, A-tailing and adapter ligation, and are then selected based on size via gel electrophoresis or magnetic beads. The TBS protocol also includes a fragment selection step via hybridization, which allows for the enrichment of specific target regions. The surviving fragments are then denaturated and treated with sodium bisulfite to deaminate unmethylated Cs into Us. PCR amplification converts Us back to Ts and the fragments are then ready to be sequenced.

Recently, a single-cell bisulfite sequencing protocol was developed, allowing for more accurate analysis of samples containing heterogeneous cell populations [117].

3.2.3 Analyzing sequencing data

All types of sequencing data require tailored software and pipelines to extract meaningful biological information. However, when performing differential expression or differential methylation analyses, the main steps are very similar, as shown in Figure 7 and described below.

Quality control. The computational analysis of RNA-Seq and Bi-Seq data starts with QC of the reads (usually encoded in fastq format) to assess the proportion of data that are of sufficiently high quality to be included in the analysis. QC is

commonly performed using software such as FastQC [118], which produces several plots to assess, among others, the distribution of quality scores at different positions in the read, the distribution of the average read quality, the read length distribution across all reads, the proportion of each nucleotide at different positions in read, and the the presence of overrepresented sequences.

Trimming and adaptor removal. Reads might contain parts of the adapter sequence at both ends; moreover, nucleotides at the extremities of a read often

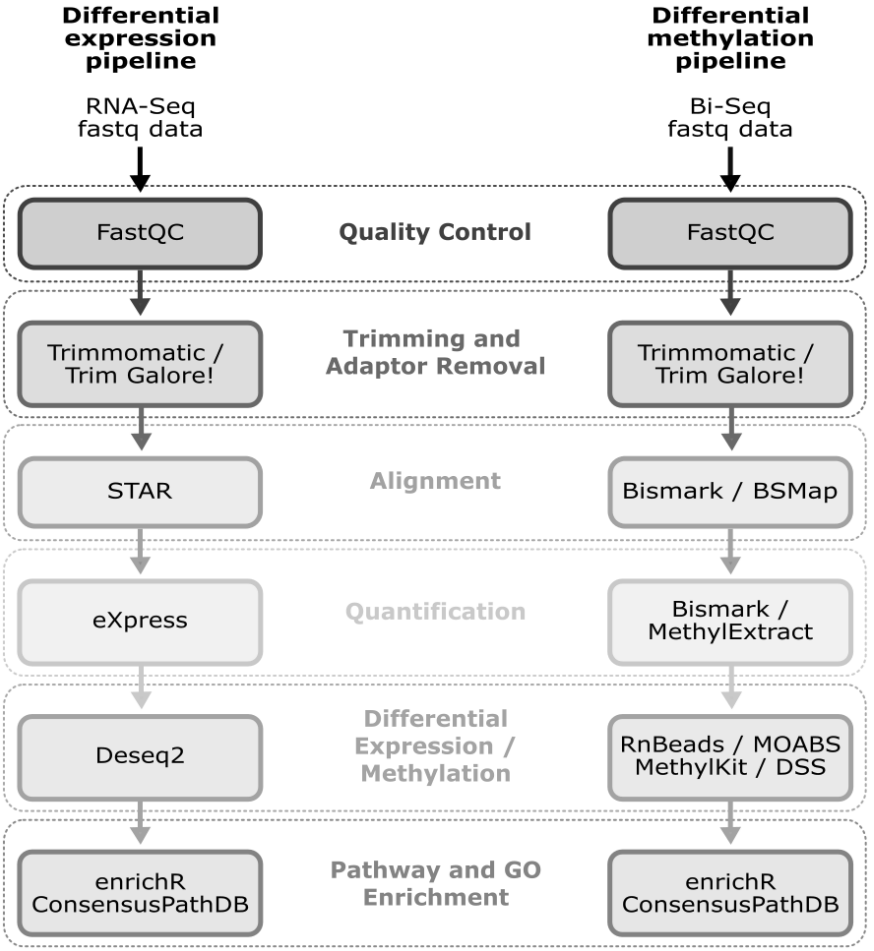

Figure 7: Standard bioinformatic pipeline describing the analysis of RNA-Seq and bisulfite sequencing data to perform differential expression / methylation analyses. Steps like quality control, trimming, and pathway enrichment analysis can be performed using the same software for both RNA-Seq and Bi-Seq data. On the other hand, tools specific for each data type are needed for alignment, quantification, and identification of differences between conditions.

display lower sequencing quality and might lead to misalignment. To minimize these sources of error, tools such as Trimmomatic [119] or Trim Galore! [120] are commonly used to trim the reads, removing adapters and low quality end sequences. Low quality reads or reads that are too short after being trimmed are discarded, and the quality of the remaining reads is assessed again.

Alignment. Reads passing these first filters are then mapped against a reference genome using aligner tools designed to address the challenges of each data type. Hence, aligners for RNA-Seq data should accounts for spliced alignments when mapping reads to the reference, while aligners for Bi-Seq data should take into account the T-to-C discrepancy caused by bisulfite conversion. STAR (Spliced Transcripts Alignment to a Reference) [121] is the most common aligner for RNA-Seq data. Even though it is memory intensive, STAR is much faster than other aligners, outperforming them in mapping speed by a factor of over 50. Bismark [122] and BSMap [123] are are frequently used to align Bi-Seq data. This class of aligner tools allows not only to map reads to a reference genome, but it also estimates whether the base mapped to a reference cytosine is a C (i.e. a methylated C) or a T (i.e. an unmethylated C). The sequence context in which the reference cytosine is embedded is also estimated, to allow user to separate CpG sites from CHG and CHH sites (where H stands either A, G, or T nucleotide).

Quantification. Once the reads are mapped, it is possible to estimate the levels of expression or methylation. Software like eXpress [124] can quantify the abundance of each gene/transcript. New software such as Kallisto [125] can perform gene/transcript quantification without first aligning the RNA-Seq reads. Common measures of gene expression are read counts and FPKM (which stands for Fragments Per Kilobase of transcript per Million mapped reads). Bi-Seq aligners such as Bismark usually include a function to extract the methylation values from aligned reads, but there are also tools such as MethylExtract [126] which can call both methylation levels as well as genetic variants. The methylation level of a specific cytosine in the genome is commonly estimated as the proportion between the number of mapped reads containing a C over the total number of reads aligned at that same position.

Differential expression/methylation. After the quantification step, different samples are compared to identify differentially expressed genes (DEGs). Tools such as the R package DESeq2 [127] are commonly used to estimate differential expression between conditions. DESeq2 takes as input the unnormalized counts and the condition to which each sample belongs to, and for every gene it estimates the log2 fold change, p-value, and adjusted p-value, which can then be used to identify genes with significant changes in abundance between conditions. Differential methylation to locate DNA methylome changes across conditions can be performed at the single

CpG level, to identify differentially methylated sites (DMSs), or at the regional level, to identify for instance differentially methylated genes (DMGs) or differentially methylated promoters (DMPs). Several bioinformatic tools are available to compute differential methylation. The R package RnBeads [128, 129] is one of the most commonly used packages, thanks to its extensive functionalities for filtering and exploring the data before computing differential methylation via Fisher's exact test. Other tools include MOABS [130], and R packages like MethylKit [131], InfiniumPurify [132], and DSS (dispersion shrinkage for sequencing data) [133].

Pathway and GO enrichment. To interpret the role that DEGs and DMGs/DMPs play, pathway analysis and gene ontology enrichment are performed using tools like enrichR [134, 135] or ConsensusPathDB [136], which automatically browse through public databases (such as KEGG [137], Reactome [138], WikiPathways [139], and gene ontology [140]) to identify the molecular and biological mechanisms affected by changes in the transcriptome.

When both RNA-Seq and Bi-Seq data are available for the same set of samples, it is also possible to integrate these two data layers to identify genes whose expression might be regulated by changes in DNA methylation patterns.

3.2.4 Public repositories of sequencing data

The huge amount of data generated by high throughput technologies created a need for building public, discipline-specific repositories, where such data can be stored and shared with the whole scientific community. Examples of such repositories include the Gene Expression Omnibus (GEO) [141], the Sequence Read Archive (SRA) [142], the International Human Epigenetic Consortium (IHEC) [143], the Cancer Cell Line Encyclopedia (CCLE) [144], and the European Genome-phenome Archive (EGA) [145].

In this thesis we used data from two public repositories: The Cancer Genome Atlas [146] and The Encyclopedia of DNA Elements [147, 148].

The Cancer Genome Atlas (TCGA) is an program aimed to characterize the molecular signatures of over 20,000 primary cancer and matched normal samples spanning 33 cancer types. Using a mix of microarray and sequencing technologies, TCGA generated genomic, epigenomic, transcriptomic, and proteomic data. These data are available in the TCGA online repository in three different formats: raw and controlled data (level 1), processed and controlled data (level 2), and segmented or interpreted data (level 3). For each sample, clinical data are also available. This allows researchers to perform pan-cancer integrative studies.

The Encyclopedia of DNA Elements (ENCODE) is a project launched in 2003

that aims to develop a comprehensive map of functional elements in the human and mouse genome. ENCODE data are publicly available and include both functional genomic data of cell lines and tissues generated through microarray and sequencing technologies, as well as functional characterization data.

3.3 Bioinformatic workflow management systems to analyze high-throughput data

The large volumes of data generated by high-throughput technologies require multi-step computational analyses in order to extract biologically meaningful information. The sequence of tasks to process a set of data is known as workflow and is commonly represented via a directed network. Workflows can be run manually, but when dealing with large datasets a fully automated approach is generally more efficient. Hence, workflow management systems (WMSs) have been developed to provide an infrastructure to design, execute, and monitor complex pipelines.

Several WMSs are currently available to perform tailored bioinformatic analyses by combining state-of-the-art computational tools. WMSs such as Galaxy [149], Taverna [150], Chipster [151], and GenePattern [152] are suitable for non-programmers, while other WMSs, like Anduril [153], Bpipe [154], Nextflow [155], and Snakemake [156], require coding skills in order to build the workflows.

A schematic representation of a bioinformatic WMS is shown in Figure 8. Bioinformatic WMSs can import multidimensional biomedical data generated by the user or available in public databases. The user can then design workflows that sequentially combine different steps in the analysis, each carried out by specific computational tools and programming languages, to process, visualize, and even integrate multiple data layers. The WMS will automatically run the pipeline using high-performance computing and generate results that can be used to draw conclusions regarding the biomedical questions under investigation.

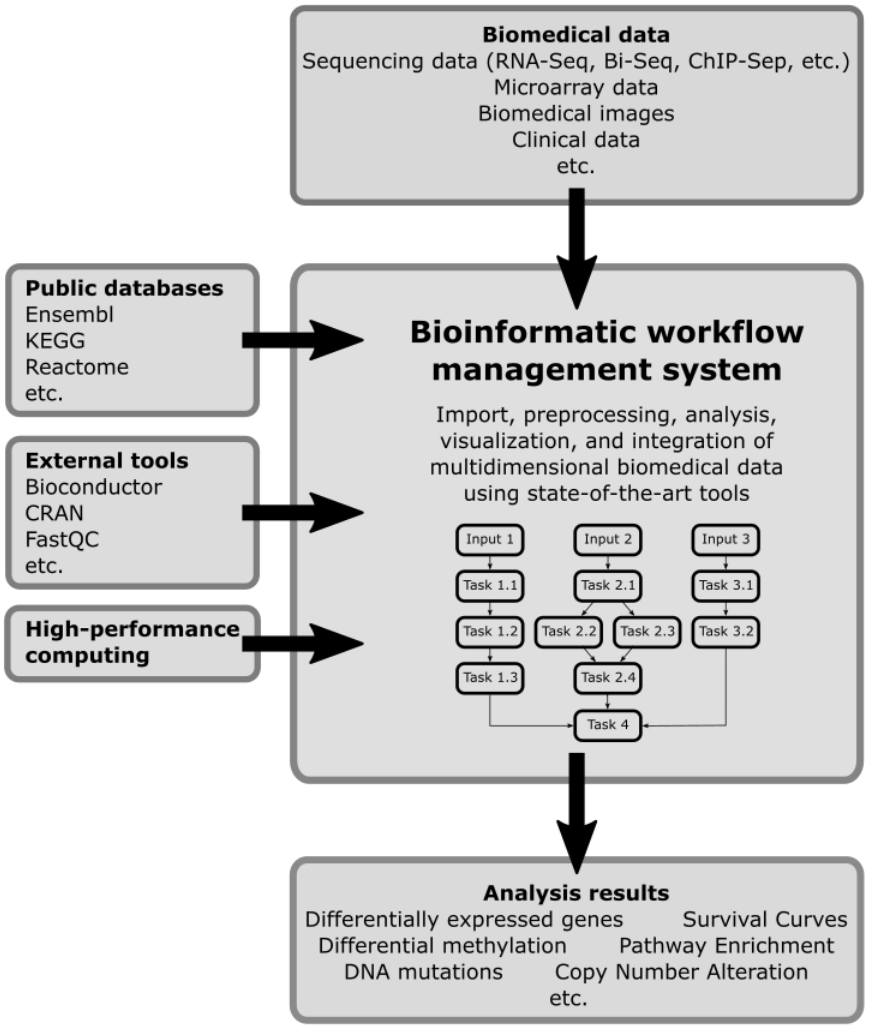

Figure 8: Schematic representation of a bioinformatic workflow management system. The WMS can import biomedical data generated by the user or available in online public repositories. The data are then processed via a workflow designed by the user, which employs external state-of-the-art bioinformatic software, but can also include user-generated scripts. The workflow is then run using high-performance computing resources to optimize the running time.

4 Aims of the study

The overall objective of this thesis is to develop *in vitro* and *in silico* methods to increase our understanding of the mechanisms behind drug resistance in cancer. To achieve this goal, we used a systems biology approach which combined different laboratory biology experiments, multivariate statistical methods and bioinformatic analyses to uncover cellular processes responsible for drug resistance. The main goals of this project are:

1. Developing a drug screening assay to discover epigenetic inhibitors able to overcome doxorubicin and rituximab resistance in DLBCL.
2. Integrating multi-omics data to identify predictive biomarkers of platinum resistance across multiple cancers.
3. Developing a computational method to estimate differential DNA methylation across cancer samples of varying tumor purity.
4. Developing a workflow management system to build complex bioinformatic pipelines to analyze and integrate multiple data types.

5 Materials and methods

In this section, the main materials and methods employed in this work are summarized.

5.1 Materials, samples and data

The patient samples and cancer cell lines used in Publications I-IV as well as in our unpublished results are summarized in Table 1, while the drugs, compounds and siRNAs used in the biological experiments of Publications I and II are listed in Table 2. Full details on the experimental designs are available in the appended manuscripts.

Publication	Samples	Technology	Source
Publication I	Oci-Ly-19, Oci-Ly-3, Riva-1, and Su-Dhl-4 DLBCL cell lines	Drug screening, RNA-Seq, WES, immunofluorescence assay	In-house
Publication II	Treatment-naive samples from 1,503 cancer patients who received platinum therapy	RNA-Seq, CGH-array, MethylArray, Clinical data	TCGA
	A549, FaDu, Kuramochi, and SW-48 cell lines	RNAi experiment to measure cisplatin response when 12 genes are knocked-down	In-house
Publication III	GM12878, K562, and HepG2 cell lines	WGBS	ENCODE
	5 HGSOC samples (2 primary and 3 relapse) and 1 white blood cell sample	TBS, RNA-Seq, WGS	In-house
	8 DLBCL samples (4 primary and 4 relapse) and 2 blood samples	RRBS, WGS	In-house
Publication IV	50 OV samples (26 long response patients and 24 short response ones)	RNA-Seq	TCGA
Unpublished data	53 HGSOC samples (27 primary, 17 interval, 9 relapse) from 14 different patients	WGBS	In-house

Table 1: Samples and cell lines used in Publication I-IV and in our unpublished analysis

Publication	Class	Compound
Publication I	BRD inhibitors	I-BET151, JQ1, OTX015, PFI-1, SGC-CBP30, UNC1215
	DNMT inhibitors	Azacitidine, Decitabine, Epigallocatechin, Lomeguatrib, Procainamide, RG108, Thioguanine
	HAT inhibitors	C646
	HDAC inhibitors	AR-42, Belinostat, CI994, CUDC-101, CUDC-907, Entinostat, Givinostat, JNJ-26481585, Mocetinostat, Panobinostat, PCI-24781, PCI-34051, Resminostat, RGFP966, Rocilinostat, Romidepsin, SB939, Tubacin, Tubastatin A, Valproic acid, Vorinostat
	HDM inhibitors	GSK J4 HCl, IOX-1, Tranylcypromine
	HMT inhibitors	EPZ005687, GSK343, Pinometostatm, Tazemetostat, UNC0638, UNC0642
	Chemotherapy	Doxorubicin
	Immunotherapy	Rituximab
Publication II	siRNAs	siRNAs targeting genes *ANXA9*, *ARC*, *C19orf33*, *ECI2*, *FBXO17*, *LGALS3BP*, *MLF1*, *MRPS21*, *PLK1*, *SEC62*, *SOX17*, *TM4SF1*, and *TSPYL5*
	Chemotherapy	Cisplatin

Table 2: Compounds and siRNAs used in Publication I-II

The unpublished data include 53 WGBS HGSOC samples extracted from 14 patients diagnosed at stage III or IV. Samples were collected during primary debulking or via laparoscopy (primary sample, p), during interval debulking (interval sample, i), or when the patient relapsed (relapse sample, r). Samples were also obtained from multiple locations, including ovary (ova), omentum (ome), peritoneum (per), intestine (meso), Fallopian tubes (tub), ascites (asc), or undefined location (tum). Samples were collected at the Department of Obstetrics and Gynecology of the Turku University Central Hospital, as part of the HERCULES project (European Union's Horizon 2020 research and innovation program under grant agreement No. 667403). Ethical approval for the study is stated in the Hospital District of Southwest Finland ethics committee statement EMTK 145/2015 22.5.2018 § 197.

5.2 Drug sensitivity screening

We used automated liquid-handling robots to implement multi-step drug screenings to investigate drug resistance in cancer.

In Publication I, we designed a drug screening to test whether pretreating DLBCL cell lines with epigenetic inhibitors can increase the response of doxorubicin and rituximab. Briefly, 60 epigenetic compounds were pre-seeded on 384-wells plates and DLBCL cells were then added. A new dose of epigenetic inhibitor therapy was administered every three days for nine days, to induce cellular reprogramming. Rituximab and doxorubicin were then administered (at fixed concentrations in the initial screening, and at varying concentrations in the validation screening). 48 hours later, viability was measured to determine whether the epigenetic treatment induced sensitization of DLBCL cells to the immunochemotherapy.

In Publication II, we tested how cisplatin response changes when genes we identified as related to platinum resistance were knocked-down. siRNAs targeting 12 biomarker genes were transfected in cancer cell lines and, after 24 hours, cisplatin was administered. Cell viability was measured after 48 hours and the effect of gene knockdown on platinum response was estimated.

5.3 Anduril workflow management system

The open source component-based workflow framework Anduril (ANalysis of Data Using Rapid Integration of aLgorithms) [153, 157] developed by the Systems Biology of Drug Resistance in Cancer group (University of Helsinki) was used to build the computational pipelines in Publications I-IV.

Pipelines implemented using the first version of Anduril were coded using the programming language AndurilScript, while Anduril 2.0 uses the Scala language to build workflows. Each pipeline runs a sequence of components, which can be created using a variety of programming languages, including R, Matlab, Java, Bash, and Python. Components are divided into bundles based on the type of data/analysis they are designed for. The pipelines in Publications I-IV combine existing components with newly developed ones, and a separate bundle of Anduril components was created to analyze dose response data from Publication I.

5.4 Differential methylation using maximum-likelihood (DMML) method

Bulk tumor samples contain a mixture of cell types that comprise both cancerous and non-cancerous cells. Thus, when comparing the methylomes of two cancer samples, it is important to account for their differing tumor purity levels. To address this issue, we developed a computational method to perform differential methylation using maximum-likelihood (DMML). The DMML algorithm estimates the purity of

two or more cancer samples, extracts the tumor-specific signal from bulk samples, and computes differential methylation more accurately. As methylation patterns might span several neighboring CpG sites, DMML models co-methylation for a user-defined number of adjacent CpG sites that are measured in the same read.

A detailed description of the DMML method can be found in Publication III. Briefly, the method uses an expectation–maximization (EM) algorithm to estimate the latent variables (in our case the DNA methylation levels in both cancer and normal cells) using the observed DNA methylation data (which comes from a mixture of tumor and non-tumor cells). DMML can take a non-tumorous control sample as input, but this is not required when estimating the purity of the samples or assessing differential methylation. Controls are only required when determining the methylome state for each cell type.

We compared the performance of DMML against other existing methods, including Fisher's exact test, MOABS, a mixture-adjusted Fisher's exact test, InfiniumPurify and DSS methods. To evaluate their performance we used simulated data generated by mixing WGBS data from three ENCODE cell lines (GM12878, K562, and HepG2). We also estimated tumor purity and differential methylation in patient-derived samples, including TBS data from HGSOCs and RRBS data for DLBCLs. We used DMML to estimate tumor purity in patient samples, and compared the obtained values to the purity estimated from genome sequencing data from the same samples.

DMML is suitable for pairwise comparison of two or a few samples, but the co-methylation modeling needs to be dropped to scale the method and allow simultaneous comparison of many (»10) samples. In this way, we were able to use DMML to extract the patient-, location-, and treatment stage-specific methylation profiles, as well as a profile from the non-cancerous cells, from samples in our unpublished HGSOC dataset.

5.5 Interactive exploration of multiple data types

Visualizing and interpreting biomedical datasets can be challenging due to the huge amounts of information they contain. To address this issue, we used the R packages shiny [158] and plotly [159] to create interactive visualizations of our datasets in Publication I. Shiny allows users to subset data and visualize different portions of the entire dataset, while plotly provides interactivity with the plot itself by hovering over it and zooming in and out.

Using these packages we developed a website (http://app.anduril.org/DLBCL_DSRT) where different data types (including sequencing data, drug response curves, and

immunofluorescence images) could be loaded and explored.

5.6 Automatic workflow to integrate multi-omics TCGA data and identify predictive biomarkers

In order to integrate multi-omics data from pan-cancer TCGA patients, we designed a pipeline that would download expression, DNA methylation and gene copy number data from the GDC data portal, and identify genes (9,976) and samples (1,503) for which all three data layers are available. DNA methylation and copy number data were then binarized to match the input format required by the R package CNAmet [160], which was then used to perform data integration and identify genes whose expression is regulated by DNA methylation or gene copy number.

5.7 Survival analysis

In biomedical research, survival analysis is an extremely useful tool to assess how certain biological and molecular features, or even certain treatment options, improve or diminish survival in a cohort of patients across time. Typical time measures are progression free survival (PFS) and overall survival (OS). PFS measures the time from diagnosis until the first recurrence, the last follow-up or death. OS measures instead the time from diagnosis until the last follow-up or death.

The survival function estimates the proportion of patients still alive at a given time and it is visualized using Kaplan-Meyer plots. The non-parametric log-rank test is commonly used to test whether the difference between two survival functions is significant. In Publication II and IV, survival analysis was performed in R via Kaplan-Meier analysis and log-rank test using the package survival [161, 162] and survival curves were visualized using the package survminer [163]. In Publication II, patients with the 20% highest and lowest expression of each gene were included in the analysis, and PFS was used as the the time variable. In Publication IV, OS was instead used to investigate the effect that certain mutations had on the survival of HGSOC patients.

6 Results

In this thesis, I present different *in vitro* and *in silico* methods designed and implemented to investigate the molecular mechanisms responsible for drug resistance in cancer. The main results from Publications I-IV, as well as some unpublished data on DNA methylation changes occurring in HGSOC patients, are summarized below.

6.1 Pretreating DLBCL cell lines with epigenetic inhibitors sensitizes them to immunochemotherapy (I)

In order to systematically assess the ability of epigenetic inhibitors to reverse R-CHOP resistance in DLBCL, we designed and implemented a multi-step drug screening. Four DLBCL cell lines (Riva-I, Su-Dhl-4, Oci-Ly-3, and Oci-Ly-19) were pretreated for 9 days using 60 different epigenetic inhibitors, and were then administered rituximab and doxorubicin (the main components of the R-CHOP regimen). Cell viability was estimated 48h later and the response of cells receiving only the epigenetic pretreatment was compared to the response of cells receiving the pretreatment followed by rituximab and doxorubicin. HDAC inhibitors induced sensitization in all cell lines, while BRD and HMT inhibitors sensitized three of the four cell lines. Oci-Ly-3 was the most responsive cell line (20/60 inhibitors inducing sensitization), followed by Su-Dhl-4 (10/60), Oci-Ly-19 (9/60) and Riva-I (3/60). It is important to note that the concentration of the epigenetic inhibitors required to induce cellular reprogramming was lower than the concentration needed to induce a cytotoxic effect, which indicates that using epigenetic inhibitors as a pretreatment instead of in combination with R-CHOP might lead to less side effects.

To validate the sensitizing effect of the most potent inhibitors, we conducted a second drug screening to estimate the synergy between the epigenetic compounds and the immunochemotherapy. In this assay, we tested varying concentrations of both the epigenetic inhibitor and the rituximab-doxorubicin combination. Ten compounds were included in this validation: belinostat, entinostat, and I-BET151 were tested in all four cell lines, while vorinostat, resminostat, givinostat, pinometostat, tazemetostat, SGC0946, and OTX015 were tested only in the cell line(s) that showed a reprogramming effect from that drug in the first screening. HDAC inhibitors (vorinostat, entinostat,resminostat, and belinostat), and HMT inhibitors (pinometostat, tazemetostat, and SGC0946) showed high synergy with rituximab and doxorubicin. Among these, entinostat and tazemetostat displayed the most potent sensitization effects. BRD inhibition showed instead lower synergy.

6.2 Molecular mechanisms responsible for epigenetic sensitization in DLBCL (I)

To investigate which genes and pathways are linked to epigenetic sensitization, we compared the trascriptomes of the four untreated DLBCL cell lines with their transcriptomes after pretreatment with entinostat, vorinostat, belinostat and tazemetostat. Interestingly, differential expression followed by pathway analysis identified disruption of DNA repair, cell cycle, cell adhesion and apoptosis as potential mechanisms behind epigenetic sensitization.

We further investigated the link between impaired DNA repair pathways and epigenetic sensitization to doxorubicin using an immunofluorescence assay. DLBCL cells treated with HDAC inhibitors (entinostat, belinostat, vorinostat) showed impaired homologous recombination (HR) and upregulation of the non-homologous end joining (NHEJ) pathway (which was expected since NHEJ is often active when HR is not functioning correctly).

When we looked into genetic mutations which might affect the efficacy of epigenetic reprogramming, we observed that the Su-Dhl-4 cell line has a missense mutation in the *EZH2* gene, which encodes an HMT enzyme. This mutation has been shown to increase sensitivity to tazemetostat [164]. Other notable mutations that might affect the efficacy of HDAC inhibitors were a truncating mutation in the *CREBBP* gene identified in Riva-I and Oci-Ly-19, and a truncating mutation in the *ARID1A* gene in Riva-I.

6.3 Identification of potential biomarkers and therapeutic targets to predict and overcome platinum resistance in multiple cancers (II)

The integration of gene expression, DNA methylation, and gene copy number data of 9,976 genes across 1,503 pan-cancer patients treated with platinum chemotherapy revealed 333 upregulated and 48 downregulated genes (FDR ≤ 0.01). Survival analysis showed that 164 out of these 381 genes have a significant association to progression free survival (PFS), and are hence promising candidate biomarkers to predict platinum response. A spreadsheet listing these genes is included in supplementary table S4 of Publication II. Among these 164 putative biomarkers we found genes involved in pathways related to drug resistance, cancer progression and metastasis, apoptosis, and immune-related mechanisms. Some of these genes, like *ARC*, were detected at high and low expression in most of the 17 cancer types included in the study. Other genes, like *SOX17*, were expressed only in one or a few

cancers. To identify the most promising pan-cancer biomarkers, we used different ranking systems to prioritize genes that had (i) both high and low expression across most cancer types, and (ii) whose expression was strongly regulated by gene copy number or DNA methylation.

To investigate whether our biomarkers could also serve as therapeutic targets to increase platinum response in multiple cancers, we performed an RNAi screening using cell lines obtained from lung (A549), head and neck (FaDu), ovarian (Kuramochi), and colorectal (SW-48) cancers. *ANXA9*, *C19orf33*, *ECI2*, *FBXO17*, *LGALS3BP*, *MLF1*, *MRPS21*, *SEC62*, *TM4SF1*, *TSPYL5* were selected as the top pan-cancer candidates, while *SOX17* was included in the screening because it is highly expressed in ovarian cancer. Cisplatin response in the untreated cell lines was compared to cisplatin response when these 12 genes were downregulated. *MRPS21* was the only gene to mildly but consistently increase cisplatin response in all cell lines. *ARC*, *ANXA9*, *C19orf33*, and *TSPLY5* showed synergy with cisplatin treatment in two cell lines. Knockdown of *MLF1* enhanced cisplatin inhibitory power in FaDu cells, while knockdown of *SOX17* improved cisplatin response in SW-48 cells but not in Kuramochi cells (the HGSOC cell line).

6.4 Estimating differential methylation in cancer samples with varying tumor purity (III)

Bulk tumor samples contain a mixture of cancerous and non-cancerous cells. Since DNA methylation patterns are cell type specific, the methylome observed by sequencing bulk samples is not the actual cancer methylome but a mixture of cancer and normal methylomes. Our DMML method allows for accurate estimation the tumor purity of a sample even without a matched control, and for performance of differential methylation while taking into account the proportion of cancer cells contained in each sample.

To test DMML's tumor purity estimation we used Monte Carlo simulations to generate tumor samples with varying levels of purity by mixing the WGBS methylation profiles of cancer cell lines (K562 and HepG2) and a non-cancerous cell line (GM12878). The accuracy of DMML purity estimation in low purity samples exceeds that of existing methods (like Fisher's exact test, MOABS, DSS, and InfiniumPurify). Moreover, when reads are long enough to contain multiple CpG sites, DMML can model co-methylation, further increasing the method's accuracy even in low coverage settings.

We then used DMML to perform differential methylation using five TBS HGSOC samples and eight RRBS DLBCL samples, to test whether the method is effective

in different cancer types as well as when using different bisulfite sequencing technologies. We compared DMML's differential methylation calls with the ones obtained from the other methods. Any two of these methods shared at least 91% of the calls in the HGSOC dataset and 83% of the calls in the DLBCL dataset, but DMML identified novel candidate differentially methylated sites even without a control sample.

We also investigated the relation between methylation and gene expression in the HGSOC dataset. Differential promoter methylation obtained with DMML showed significant anticorrelation with differential expression. The anticorrelation between methylation and expression was stronger when using differentially methylated sites identified by DMML than those identified by other methods, suggesting once again that DMML can more accurately recover the cancer methylation signal in bulk samples.

6.5 Deconvoluted DNA methylation from 53 HGSOC samples (Unpublished data)

Heterogeneity in a tumor sample occurs at different levels, including inter-patient heterogeneity, clonal evolution of the metastatic lesions driven by the microenvironment, clonal evolution due to treatment exposure, and presence of non-cancerous cells within the tumor mass. When analyzing DNA methylation data obtained from bulk tumor samples, it is important to remember that the measured signal is mixture of all these heterogeneous components.

In order to estimate how much patient variability, tumor location, sampling stage and non-cancerous methylomes contribute to the measured methylation signal, we used a modified version of the DMML method to deconvolute the methylome of each sample into these four signals. According to our results shown in Figure 9, patient-specific variability contributes the most to the methylation signal observed in our samples. This is not surprising since DNA methylation patterns are affected by environment and lifestyle. The sample location also affects the methylation landscape quite considerably, especially in the five samples obtained during the interval surgery. This could indicate that the selective pressure of the tumor microenvironment at the metastatic site has an effect on the DNA methylation patterns of the cancer cells.

6.6 Anduril 2 workflow framework (IV)

The use of Anduril 2 as a WMS to implement complex computational analyses presents several advantages. The downloadable version of Anduril 2 already in-

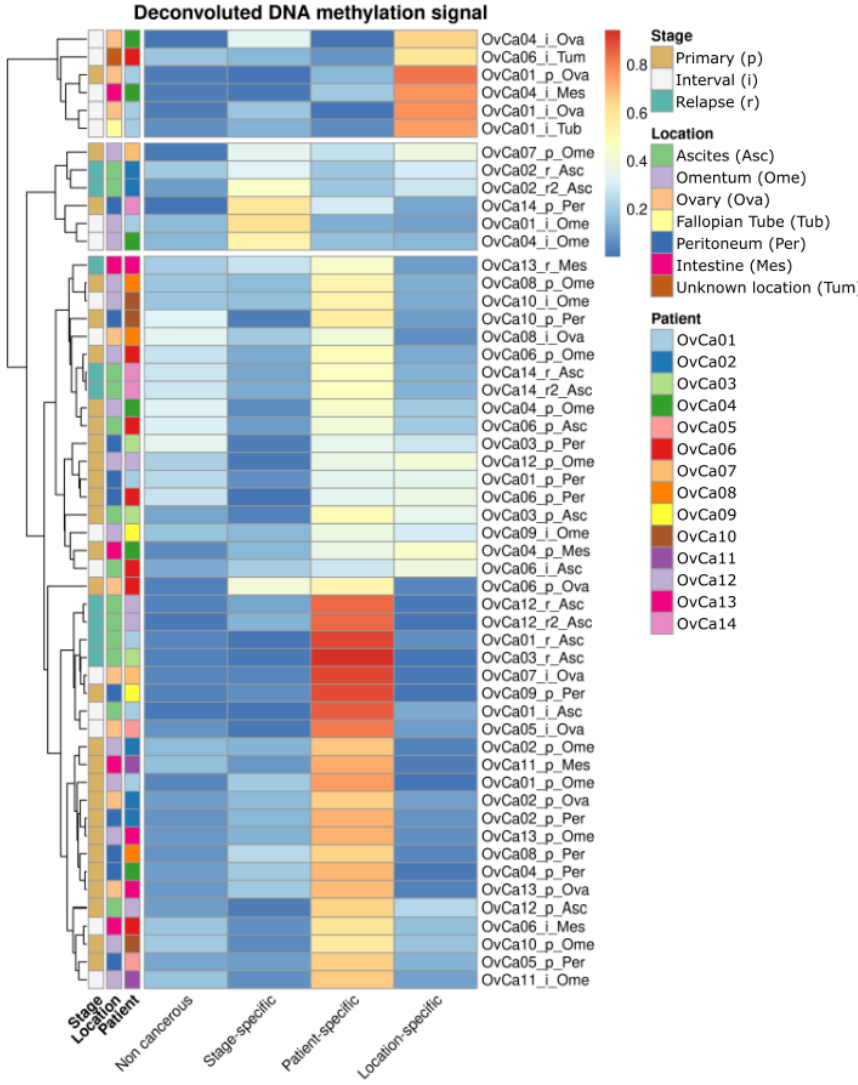

Figure 9: Heatmap showing the contribution of deconvoluted DNA methylation signals from 53 HGSOC samples. For each sample we estimated the proportion of the observed methylation signal that is due to inter-patient variability (patient-specific component), microenvironment-driven evolution (location-specific component), treatment stage at which the sample was obtained (stage-specific component), and proportion of the signal due to non-cancerous cells present in the sample (non-cancerous component).

cludes over 400 fully-documented components and functions, suitable for analyzing all main NGS data types. New components can be easily developed by users, providing a flexible and modular approach to design tailored workflows. Moreover, component dependencies are automatically detected and independent parts are

parallelized to reduce processing time. If the workflow execution is interrupted or the user wants to change parts of the script, only the modified/unprocessed portion of the analysis needs to be re-run. This is particularly useful in bioinformatic analyses processing large datasets with run-times of several days.

7 Discussion

Overcoming drug resistance in cancer is one of the major health challenges of the 21st century. Modern high throughput screening and sequencing technologies allow us to explore the molecular mechanisms of cancer cells at an unprecedented level of detail. This, in turn, aids the search for effective treatment strategies to overcome drug resistance, as well as the discovery of biomarkers able to predict patients' response to specific treatments. Moreover, the increased availability of large cancer datasets in public repositories provides scientists with invaluable resources needed to develop novel methods and software to analyze these data, giving rise to the field of computational oncology [165].

The work presented in this thesis exploits high throughput technologies to develop *in vitro* and *in silico* methods to investigate R-CHOP resistance in DLBCL, as well as platinum resistance across multiple cancers, and more specifically in HGSOC.

In Publication I, we developed a drug screening method to measure the effect of non-simultaneous drug combinations on cancer cell lines. We used this method to test the ability of 60 epigenetic inhibitors to sensitize DLBCL cell lines to rituximab and doxorubicin. Since the experiment was conducted *in vitro*, we could only investigate how epigenetic reprogramming affects the cytotoxic effect of rituximab [75] and not its ability to induce an immune response towards the cancer cells. HDAC and HMT inhibitors were the most promising classes of compounds to sensitize DLBCL cells to the toxic effect of rituximab and doxorubicin. The inhibitor concentration required to achieve sensitization was lower than then one required to produce a cytotoxic effect, suggesting that the use of epigenetic inhibitors as a pretreatment before R-CHOP administration in clinical use could result in less side effects than if they were used simultaneously. When investigating the mechanisms that might be involved in the episensitization process, we uncovered that HDAC inhibitors are able to dysregulate the homologous recombination (HR) pathway, which in turn increases the efficacy of doxorubicin. This finding is in line with other studies showing a link between HDAC inhibition and downregulation of HR[166, 167], suggesting that HDAC inhibitors could be effective sensitizers for DNA damaging treatements, including radiotherapy and chemotherapeutics like platinum-based drugs.

The use of epigenetic inhibitors, alone or in combination with immunochemotherapies, has already shown promising results in several clinical trials [168]. However, undesired side effects on non-cancerous tissues need to be minimized, and targeted nanoscale delivery systems promise to improve the effectiveness of epigenetic treatments in cancer [169].

In Publication II, we took advantage of public cancer data from The Cancer Genome Atlas (TCGA) repository to investigate platinum resistance in a pan-cancer context. We hypothesized that combining evidence from three different data levels (expression, copy-number and DNA methylation) would enable us to overcome spurious findings and identify genes related to platinum resistance. We performed multi-omics data integration followed by survival analysis to identify genes whose expression is related to platinum response. We identified 164 putative biomarkers predictive of platinum response. Further investigation of these candidates could lead to the development of a routine test to select which patients should receive platinum treatment and which patients should be administered a different therapeutic regimen. We also conducted an RNAi experiment to test whether the most promising candidate biomarkers could also serve as therapeutic targets to overcome platinum resistance, and we were able to find an association between the knockdown of the gene encoding the mitochondrial ribosomal protein *MRPS21* and an increased response to cisplatin in four different cancers. The exact function of MRPS21 is still poorly understood, however other mitochondrial ribosomal proteins have been reported to affect apoptotic mechanisms [170].

In Publication III, we developed a new bioinformatic software (DMML) to correctly estimate differential methylation in cancer samples with varying tumor purity. DMML can be used to analyze different types of bisulfite sequencing data (including WGBS, RRBS, and TBS), outperforms existing alternatives, produces robust results with and without non-cancerous controls, does not require prior purity estimates, and can be used to to compute tumor purity and differential methylation in any cancer type. Since DNA methylation has been shown to be a contributing factor in the development of drug resistance, understanding the link between methylome changes and drug response can help us to develop strategies to overcome treatment failure. We also used a modified version of the DMML algorithm to estimate the origin of inter-sample variability, as shown in our unpublished results. According to our analysis, the differences in methylation patterns observed in our 53 HGSOC samples are patient-specific, likely due to adaptation to diverse environmental stimuli accumulated throughout the patient's life, while changes driven by treatment stage and sample location (*i.e.* the pressure to adapt to the microenvironment) contribute less to the observed methylation signal.

In publication IV, we develop Anduril 2, a workflow management system able handle the processing of big data obtained from high-throughput biology experiments. This WMS has been used to implement the data analysis pipelines of publication I and II, while the software described in publication III has been included as a standard component in Anduril 2 and can easily be included in future pipelines.

Taken together, the methods presented in this thesis provide new tools to investigate

the molecular mechanisms behind cancer progression and drug resistance. Moreover, these methods are flexible and can be applied to biomedical investigation beyond the cancer research field.

With new and continuously emerging high-throughput technologies, like single-cell and third generation sequencing, computational oncology is a fast-evolving field which will gain more and more of a central role in the fight against cancer in the coming years. The integration of multi-level cancer data holds the key to unlocking the complexity of drug resistance, generating new insights to improve cancer treatment strategies, and even opening the doors to the era of precision oncology [171].

Acknowledgements

Research is always the result of teamwork, and the studies described in this thesis would not have been possible without the help and support of many people. First of all, I want to thank professor Sampsa Hautaniemi for the guidance and mentorship provided throughout my PhD. Thank you for the opportunity to dive into the world of epigenetics, and for the chance to work on both in vitro and in silico projects, which really helped me broaden my understanding of the scientific method. Thank you also for allowing me to explore the academic environment beyond the research field, and for giving me the opportunity to be a teaching assistant in your courses.

I am grateful to the members of my thesis committee, professor Merja Heinäniemi and doctor Miina Ollikainen, for their time and their useful feedback and guidance during our yearly meetings. I also thank my thesis pre-examiners, doctor Kirsi Granberg and professor Tapio Pahikkala, for the time spent revising this work and for the helpful suggestions on how to improve it.

I am very thankful to the Doctoral Programme in Biomedicine (DPBM) and the Doctoral School in Health Sciences (DSHealth) for the financial and academic support provided during my doctoral studies. In particular, I want to thank Alma, Anita, Eeva, Nina, and Hannu for their support to the PhD student community. I am also grateful for the funding provided by the Academy of Finland, the Sigrid Jusélius Foundation, and the Cancer Foundation Finland, as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 667403 for HERCULES), which were instrumental to conduct this work.

This work was carried out in the Systems Biology of Drug Resistance in Cancer Laboratory at the Faculty of Medicine, University of Helsinki during 2014-2021. I am grateful to all my colleagues at the lab for creating a very collaborative research environment. In particular, I want to thank Julia for sharing the long hours and all the ups and downs of the journey that led to Publication I. Ville, thank you for being a very patient teacher and for your help with all sorts of coding issues. Antti, thank you for being the statistics wizard of the lab! Rainer, thank you for always reminding me that biology doesn't function in binary mode, for helping me to make sense of the long gene and pathway lists, and for hosting the annual Salmari competition: making salmiakkikossu is a skill I wasn't expecting to acquire during my PhD, but which might prove useful while living in Finland. Tiia and Karen, thank you for all the invaluable support you provide, the lab could not function without you! A big thank you also to the past and current members of the lab, including Alejandra, Alexandra, Amjad, Chengyu, Emilia, Ingrid, Kaiyang, Kari, Katherine, Kristian, Jaana, Javier, Juha, Lilli, Mai, Melanie, Mikko, Oskari,

Pekka, Ping, Riku, Sanaz, Valeria, Veli-Matti, and Yilin. Thank you for all the coding camps, the Anduril days, the Monday cookie breaks, and the interesting discussions.

I am really thankful to all the research collaborators I had the opportunity to work with and learn from throughout my PhD. I thank professor Sirpa Leppä and doctor Suvi-Katri Leivonen for the interesting collaboration in the epigenetic reprogramming project. I also want to thank Anne Aarnio and Marika Tuukkanen for teaching me the basics of wet lab work. Professor Krister Wennerberg, Laura Turunen, Jani Saarela, Swapnil Potdar, and the team at the High Throughput Biomedicine Unit in FIMM, thank you for your help in implementing the drug screenings and in analyzing and interpreting the results. I am grateful to professor Liisa Kauppi and doctor Manuela Tumiati for their contribution in the epigenetic reprogramming study. A big thank you goes also to doctor Laura Lehtinen for helping me with the RNAi experiment. To professor Benno Schwikowski, thank you for welcoming me in your lab during my visit at Institute Pasteur, and for the many interesting discussions. I also want to thank all the people in HERCULES and ONCOSYS for providing a very stimulating research environment.

The research community at the University of Helsinki is very dynamic, and I am really lucky to have been able to collaborate with many scientists interested in bridging the gap between science and society. I want to thank Shishir, Elina, Mridul, Tuomo, Kul, Siggi, Hannu, and the whole TEDxHelsinkiUniversity team. I had so much fun working with you to bring this event to life, and I found many friends in the process. A big thank you also to all the past and current members of The Science Basement, and in particular to Lea, Ekaterina, Erika, Stephany, Katja, Eleanna, Alok, Susanna, and Vesa, who have really been driving this group! You are all incredibly creative and resourceful, and I always have a great time experimenting with you different ways of communicating science. I am also grateful to the members of the DPBM student council, for bringing our community closer.

Throughout my PhD, I really enjoyed the opportunity to design new courses and I had a great time working with experts that were ready to share their knowledge. I thank all the students I had the pleasure to work with, as well as Miina Ollikainen, Nina Kaminen-Ahola, Vanessa Fuller, Brendan Battersby, Vincenzo Cerullo, Oana Velcu-Laitinen, Tarja Kahkonen, Asta Pirskanen, Jussi Heinonkoski, Hannu Sariola, and all other people I collaborated with in different teaching activities. A special thank you goes to Vincenzo and Vanessa, who have been friends and mentors in these last few years.

A huge thank you to Jonna Katajisto and Olli Silvennoinen, for taking a chance on me and welcoming me into the HiLIFE team. I also want to thank Pia Runeberg,

for helping me get started with Y Science.

I definitely would not have survived this roller-coaster without good friends, always there to share the happy times as well as the hard one. To Ceci, Greta, Ausi, Pepe, Davide, Sid, Leo, and Masha, thank you for being a constant source of joy and laughter, for pampering me every time I come back to Italy, and for always being close despite the distance. To my Italians here in Helsinki, Elisabetta, Rossella, Sara, Manlio, and Stephany, thank you for always being ready to discuss and cook Italian food, and for helping me be a little less homesick (especially during this pandemic). To Evisa, Ville, and Lea, thank you for keeping me sane (as much as my three neurons allowed) and for listening to my endless talks (especially those about epigenetics). To Tiia, thank you for understanding my color-coordination OCD and for getting me to do at least a bit of exercise during my PhD years. I also want to thank Dima for being a constant source of support and a lunch buddy since our MSc studies up to the writing of this thesis. Mårten, thank you for being my bridge to the Swedish-speaking community in Helsinki, for the very colorful risottos, and for bringing Fabian into my life. To Anja and Antonia Maria, thank you for sharing the best flat in Helsinki, and for all the awesome parties. Axel, thank you for allowing me to join you and Fabian in your Survivor evenings. To all the other wonderful friends I found here in Helsinki, thank you for all the hugs and the fun times.

Together with my friends, my family has also been a source of love and endless support. I want to thank Outi, Kasper, Jaffi, Mario, Robin, Vicky, and Liam for welcoming me among the Nybergs. To my whole family in Italy (which also includes the Carollos and the El Filaouis), thank you for being big, warm, and cuddly.

I thank my parents, Mariolina and Ottavio, for always being just a phone call away, for encouraging me to jump, being there to catch me if I fall, and helping me get back up again. Thank you for all the opportunities you gave me, and for the loving family you built. I am grateful to you in more ways than I can count. Lucy, my sister, my friend, my travel companion. I am proud of the woman you have become, and I know we can always count on each other.

Fabian, thank you for being with me throughout this whole roller-coaster, for being kind, patient, positive, brave, and loving, for letting me sing to my heart's content, and for pushing me to explore the world in unexpected directions.

Chiara Facciotto
Helsinki, 2021

References

Page(s)

[1] National Cancer Institute. (2015) What is cancer? (`https://www.cancer.gov/about-cancer/understanding/what-is-cancer#types`). 1

[2] World Health Organization. (2018) Latest global cancer data (`https://www.who.int/cancer/PRGlobocanFinal.pdf`). 1

[3] World Health Organization. (2018) Cancer - key facts (`https://www.who.int/news-room/fact-sheets/detail/cancer`). 1

[4] You, J. S & Jones, P. A. (2012) Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22**, 9–20. 1

[5] Muir, P, Li, S, Lou, S, Wang, D, Spakowicz, D. J, Salichos, L, Zhang, J, Weinstock, G. M, Isaacs, F, Rozowsky, J, & Gerstein, M. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* **17**, 53. 1

[6] Friedberg, J. W. (2011) Relapsed/refractory diffuse large B-cell lymphoma. *Hematology Am Soc Hematol Educ Program* **2011**, 498–505. 1, 13

[7] Torre, L. A, Trabert, B, DeSantis, C. E, Miller, K. D, Samimi, G, Runowicz, C. D, Gaudet, M. M, Jemal, A, & Siegel, R. L. (2018) Ovarian cancer statistics, 2018. *CA Cancer J Clin* **68**, 284–296. 1

[8] Aunan, J. R, Cho, W. C, & Søreide, K. (2017) The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis* **8**, 628–642. 3

[9] Wu, S, Zhu, W, Thompson, P, & Hannun, Y. A. (2018) Evaluating intrinsic and non-intrinsic cancer risk factors. *Nat Commun* **9**, 3490. 3

[10] Renkawitz, R. (2006) *Transcription Factors and Regulation of Gene Expression*. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 1886–1890. 3

[11] Jaenisch, R & Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245–254. 4

[12] Gibney, E. R & Nolan, C. M. (2010) Epigenetics and gene expression. *Heredity (Edinb)* **105**, 4–13. 4

[13] Hannon, G. J. (2002) RNA interference. *Nature* **418**, 244–251. 4, 6

[14] Lyko, F. (2018) The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet* **19**, 81–92. 4

[15] Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484–492. 4

[16] Illingworth, R. S & Bird, A. P. (2009) CpG islands–'a rough guide'. *FEBS Lett* **583**, 1713–1720. 4

[17] Yin, Y, Morgunova, E, Jolma, A, Kaasinen, E, Sahu, B, Khund-Sayeed, S, Das, P. K, Kivioja, T, Dave, K, Zhong, F, Nitta, K. R, Taipale, M, Popov, A, Ginno, P. A, Domcke, S, Yan, J, Schübeler, D, Vinson, C, & Taipale, J. (2017) Impact of cytosine methylation on DNA 4 binding specificities of human transcription factors. *Science* **356**.

[18] Medvedeva, Y. A, Khamis, A. M, Kulakovskiy, I. V, Ba-Alawi, W, Bhuyan, M. S, Kawaji, H, Lassmann, T, Harbers, M, Forrest, A. R, & Bajic, V. B. (2014) Effects of cytosine 4 methylation on transcription factor binding sites. *BMC Genomics* **15**, 119.

[19] Shukla, S, Kavak, E, Gregory, M, Imashimizu, M, Shutinoski, B, Kashlev, M, Oberdoerffer, P, Sandberg, R, & Oberdoerffer, S. (2011) CTCF-promoted RNA polymerase II pausing 4 links DNA methylation to splicing. *Nature* **479**, 74–79.

[20] Mariño-Ramírez, L, Kann, M. G, Shoemaker, B. A, & Landsman, D. (2005) Histone 4 structure and nucleosome stability. *Expert Rev Proteomics* **2**, 719–729.

[21] Lu, D. (2013) Epigenetic modification enzymes: catalytic mechanisms and inhibitors. *Acta* 4 *Pharmaceutica Sinica B* **3**, 141 – 149.

[22] Schuettengruber, B, Chourrout, D, Vervoort, M, Leblanc, B, & Cavalli, G. (2007) Genome 6 regulation by polycomb and trithorax proteins. *Cell* **128**, 735–745.

[23] Spicuglia, S & Vanhille, L. (2012) Chromatin signatures of active enhancers. *Nucleus* **3**, 6 126–131.

[24] Huntzinger, E & Izaurralde, E. (2011) Gene silencing by microRNAs: contributions of 6 translational repression and mRNA decay. *Nat Rev Genet* **12**, 99–110.

[25] Mohr, S. E, Smith, J. A, Shamu, C. E, Neum?ller, R. A, & Perrimon, N. (2014) RNAi screening comes of age: improved techniques and complementary approaches. *Nat. Rev.* 6, 20 *Mol. Cell Biol.* **15**, 591–600.

[26] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 7 646–674.

[27] Garber, J. E & Offit, K. (2005) Hereditary cancer predisposition syndromes. *J Clin Oncol* 7 **23**, 276–292.

[28] Negrini, S, Gorgoulis, V. G, & Halazonetis, T. D. (2010) Genomic instability–an evolving 8 hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220–228.

[29] Loewe, L & Hill, W. G. (2010) The population genetics of mutations: good, bad and 8 indifferent. *Philos Trans R Soc Lond B Biol Sci* **365**, 1153–1167.

[30] Martínez-Jiménez, F, Muiños, F, Sentís, I, Deu-Pons, J, Reyes-Salazar, I, Arnedo-Pac, C, Mularoni, L, Pich, O, Bonet, J, Kranas, H, Gonzalez-Perez, A, & Lopez-Bigas, N. (2020) A 8 compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555–572.

[31] Huang, D, Duan, H, Huang, H, Tong, X, Han, Y, Ru, G, Qu, L, Shou, C, & Zhao, Z. (2016) Cisplatin resistance in gastric cancer cells is associated with HER2 upregulation-induced 8 epithelial-mesenchymal transition. *Sci Rep* **6**, 20502.

[32] Gorski, J. W, Ueland, F. R, & Kolesar, J. M. (2020) CCNE1 Amplification as a Predictive Biomarker of Chemotherapy Resistance in Epithelial Ovarian Cancer. *Diagnostics (Basel)* **10**. 8

[33] Ilango, S, Paital, B, Jayachandran, P, Padma, P. R, & Nirmaladevi, R. (2020) Epigenetic alterations in cancer. *Front Biosci (Landmark Ed)* **25**, 1058–1109. 8

[34] Kulis, M & Esteller, M. (2010) DNA methylation and cancer. *Adv Genet* **70**, 27–56. 8

[35] Esteller, M, Silva, J. M, Dominguez, G, Bonilla, F, Matias-Guiu, X, Lerma, E, Bussaglia, E, Prat, J, Harkes, I. C, Repasky, E. A, Gabrielson, E, Schutte, M, Baylin, S. B, & Herman, J. G. (2000) Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* **92**, 564–569. 8

[36] Zhao, Z & Shilatifard, A. (2019) Epigenetic modifications of histones in cancer. *Genome Biol* **20**, 245. 8

[37] Li, B & Chng, W. J. (2019) EZH2 abnormalities in lymphoid malignancies: underlying mechanisms and therapeutic implications. *J Hematol Oncol* **12**, 118. 8, 10

[38] Huffman, D. M, Grizzle, W. E, Bamman, M. M, Kim, J. S, Eltoum, I. A, Elgavish, A, & Nagy, T. R. (2007) SIRT1 is significantly elevated in mouse and human prostate cancer. *Cancer Res* **67**, 6612–6618. 8

[39] Farman, F. U, Iqbal, M, Azam, M, & Saeed, M. (2018) Nucleosomes positioning around transcriptional start site of tumor suppressor (Rbl2/p130) gene in breast cancer. *Mol Biol Rep* **45**, 185–194. 8

[40] Anastasiadou, E, Jacob, L. S, & Slack, F. J. (2018) Non-coding RNA networks in cancer. *Nat Rev Cancer* **18**, 5–18. 8

[41] McGranahan, N & Swanton, C. (2017) Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628. 9

[42] Hinohara, K & Polyak, K. (2019) Intratumoral Heterogeneity: More Than Just Mutations. *Trends Cell Biol* **29**, 569–579. 9

[43] Baghban, R, Roshangar, L, Jahanban-Esfahlan, R, Seidi, K, Ebrahimi-Kalan, A, Jaymand, M, Kolahian, S, Javaheri, T, & Zare, P. (2020) Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun Signal* **18**, 59. 9

[44] Lugano, R, Ramachandran, M, & Dimberg, A. (2020) Tumor angiogenesis: causes, consequences, challenges and opportunities. *Cell Mol Life Sci* **77**, 1745–1770. 9

[45] Whiteside, T. L & Parmiani, G. (1994) Tumor-infiltrating lymphocytes: their phenotype, functions and clinical use. *Cancer Immunol Immunother* **39**, 15–21. 9

[46] Sahai, E, Astsaturov, I, Cukierman, E, DeNardo, D. G, Egeblad, M, Evans, R. M, Fearon, D, Greten, F. R, Hingorani, S. R, Hunter, T, Hynes, R. O, Jain, R. K, Janowitz, T, Jorgensen, C, Kimmelman, A. C, Kolonin, M. G, Maki, R. G, Powers, R. S, Puré, E, Ramirez, D. C, Scherz-Shouval, R, Sherman, M. H, Stewart, S, Tlsty, T. D, Tuveson, D. A, Watt, F. M, Weaver, V, Weeraratna, A. T, & Werb, Z. (2020) A framework for advancing our understanding of 9 cancer-associated fibroblasts. *Nat Rev Cancer* **20**, 174–186.

[47] Chaplin, D. D. (2010) Overview of the immune response. *J Allergy Clin Immunol* **125**, 10 3–23.

[48] Yuen, G. J, Demissie, E, & Pillai, S. (2016) B lymphocytes and cancer: a love-hate 10 relationship. *Trends Cancer* **2**, 747–757.

[49] Armitage, J. O, Gascoyne, R. D, Lunning, M. A, & Cavalli, F. (2017) Non-Hodgkin 10 lymphoma. *Lancet* **390**, 298–310.

[50] Bray, F, Ferlay, J, Soerjomataram, I, Siegel, R. L, Torre, L. A, & Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 10 cancers in 185 countries. *CA Cancer J Clin* **68**, 394–424.

[51] Li, S, Young, K. H, & Medeiros, L. J. (2018) Diffuse large b-cell lymphoma. *Pathology* **50**, 10 74 – 87. 50th anniversary review issue.

[52] Smith, A, Howell, D, Patmore, R, Jack, A, & Roman, E. (2011) Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network. 10 *Br J Cancer* **105**, 1684–1692.

[53] Jiang, Y & Melnick, A. (2015) The epigenetic basis of diffuse large B-cell lymphoma. 10 *Semin Hematol* **52**, 86–96.

[54] Momenimovahed, Z, Tiznobaik, A, Taheri, S, & Salehiniya, H. (2019) Ovarian cancer in 11 the world: epidemiology and risk factors. *Int J Womens Health* **11**, 287–299.

[55] Zhang, S, Dolgalev, I, Zhang, T, Ran, H, Levine, D. A, & Neel, B. G. (2019) Both fallopian tube and ovarian surface epithelium are cells-of-origin for high-grade serous 11 ovarian carcinoma. *Nat Commun* **10**, 5367.

[56] Bell, D, Berchuck, A, Birrer, M, Chien, J, Cramer, D, Dao, F, Dhir, R, DiSaia, P, Gabra, H, Glenn, P, Godwin, A, Gross, J, Hartmann, L, Huang, M, Huntsman, D, Iacocca, M, Imielinski, M, Kalloger, S, Karlan, B, Levine, D, Mills, G, Morrison, C, Mutch, D, Olvera, N, Orsulic, S, Park, K, Petrelli, N, Rabeno, B, Rader, J, Sikic, B, Smith-McCune, K, Sood, A, Bowtell, D, Penny, R, Testa, J, Chang, K, Dinh, H, Drummond, J, Fowler, G, Gunaratne, P, Hawes, A, Kovar, C, Lewis, L, et al. (2011) Integrated genomic analyses of ovarian 11 carcinoma. *Nature* **474**, 609–615.

[57] Lisio, M. A, Fu, L, Goyeneche, A, Gao, Z. H, & Telleria, C. (2019) High-Grade Serous 11, 14 Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *Int J Mol Sci* **20**.

[58] Buys, S. S, Partridge, E, Black, A, Johnson, C. C, Lamerato, L, Isaacs, C, Reding, D. J, Greenlee, R. T, Yokochi, L. A, Kessel, B, Crawford, E. D, Church, T. R, Andriole, G. L, Weissfeld, J. L, Fouad, M. N, Chia, D, O'Brien, B, Ragard, L. R, Clapp, J. D, Rathmell, J. M, Riley, T. L, Hartge, P, Pinsky, P. F, Zhu, C. S, Izmirlian, G, et al. (2011) Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. *JAMA* **305**, 2295–2303. 11

[59] Narod, S. (2016) Can advanced-stage ovarian cancer be cured? *Nat Rev Clin Oncol* **13**, 255–261. 11

[60] Wyld, L, Audisio, R. A, & Poston, G. J. (2015) The evolution of cancer surgery and future perspectives. *Nat Rev Clin Oncol* **12**, 115–124. 11

[61] Chabner, B. A & Roberts, T. G. (2005) Timeline: Chemotherapy and the war on cancer. *Nat Rev Cancer* **5**, 65–72. 12

[62] Weiner, L. M, Dhodapkar, M. V, & Ferrone, S. (2009) Monoclonal antibodies for cancer immunotherapy. *Lancet* **373**, 1033–1040. 12

[63] Robert, C. (2020) A decade of immune-checkpoint inhibitors in cancer therapy. *Nat Commun* **11**, 3801. 12

[64] Berraondo, P, Sanmamed, M. F, Ochoa, M. C, Etxeberria, I, Aznar, M. A, Pérez-Gracia, J. L, Rodríguez-Ruiz, M. E, Ponz-Sarvise, M, Castañón, E, & Melero, I. (2019) Cytokines in clinical cancer immunotherapy. *Br J Cancer* **120**, 6–15. 12

[65] Hollingsworth, R. E & Jansen, K. (2019) Turning the corner on therapeutic cancer vaccines. *NPJ Vaccines* **4**, 7. 12

[66] Mohanty, R, Chowdhury, C. R, Arega, S, Sen, P, Ganguly, P, & Ganguly, N. (2019) CAR T cell therapy: A new era for cancer treatment (Review). *Oncol Rep* **42**, 2183–2195. 12

[67] Nowakowski, G, Frontzek, F, & Schmitz, N. (2019) in *Aggressive Lymphomas. Hematologic Malignancies*, eds. Lenz, G & Salles, G. (Springer, Cham), pp. 145–155. 13

[68] Pierpont, T. M, Limper, C. B, & Richards, K. L. (2018) Past, Present, and Future of Rituximab-The World's First Oncology Monoclonal Antibody Therapy. *Front Oncol* **8**, 163. 13

[69] Dotan, E, Aggarwal, C, & Smith, M. R. (2010) Impact of Rituximab (Rituxan) on the Treatment of B-Cell Non-Hodgkin's Lymphoma. *P T* **35**, 148–157. 13

[70] Lee, L, Crump, M, Khor, S, Hoch, J. S, Luo, J, Bremner, K, Krahn, M, & Hodgson, D. C. (2012) Impact of rituximab on treatment outcomes of patients with diffuse large b-cell lymphoma: a population-based analysis. *Br J Haematol* **158**, 481–488. 13

[71] Vose, J. M, Link, B. K, Grossbard, M. L, Czuczman, M, Grillo-Lopez, A, Gilman, P, Lowe, A, Kunkel, L. A, & Fisher, R. I. (2001) Phase II study of rituximab in combination with chop chemotherapy in patients with previously untreated, aggressive non-Hodgkin's lymphoma. *J Clin Oncol* **19**, 389–397. 13

[72] Sehn, L. H, Donaldson, J, Chhanabhai, M, Fitzgerald, C, Gill, K, Klasa, R, MacPherson, N, O'Reilly, S, Spinelli, J. J, Sutherland, J, Wilson, K. S, Gascoyne, R. D, & Connors, J. M. (2005) Introduction of combined CHOP plus rituximab therapy dramatically improved outcome of diffuse large B-cell lymphoma in British Columbia. *J Clin Oncol* **23**, 5027–5033.
13

[73] Pfreundschuh, M, Trümper, L, Osterborg, A, Pettengell, R, Trneny, M, Imrie, K, Ma, D, Gill, D, Walewski, J, Zinzani, P. L, Stahel, R, Kvaloy, S, Shpilberg, O, Jaeger, U, Hansen, M, Lehtinen, T, López-Guillermo, A, Corrado, C, Scheliga, A, Milpied, N, Mendila, M, Rashford, M, Kuhnt, E, & Loeffler, M. (2006) CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera International Trial (MInT)
13 Group. *Lancet Oncol* **7**, 379–391.

13 [74] Weiner, G. J. (2010) Rituximab: mechanism of action. *Semin Hematol* **47**, 115–123.

[75] Singh, V, Gupta, D, Arora, R, Tripathi, R. P, Almasan, A, & Macklis, R. M. (2014) Surface levels of CD20 determine anti-CD20 antibodies mediated cell death in vitro. *PLoS One* **9**,
13, 41 e111113.

[76] Tacar, O, Sriamornsak, P, & Dass, C. R. (2013) Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. *J Pharm Pharmacol* **65**, 157–
13 170.

[77] Thorn, C. F, Oshiro, C, Marsh, S, Hernandez-Boussard, T, McLeod, H, Klein, T. E, & Altman, R. B. (2011) Doxorubicin pathways: pharmacodynamics and adverse effects.
13 *Pharmacogenet Genomics* **21**, 440–446.

[78] Lisio, M. A, Fu, L, Goyeneche, A, Gao, Z. H, & Telleria, C. (2019) High-Grade Serous
13 Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *Int J Mol Sci* **20**.

[79] Vergote, I, Tropé, C. G, Amant, F, Kristensen, G. B, Ehlen, T, Johnson, N, Verheijen, R. H, van der Burg, M. E, Lacave, A. J, Panici, P. B, Kenter, G. G, Casado, A, Mendiola, C, Coens, C, Verleye, L, Stuart, G. C, Pecorelli, S, & Reed, N. S. (2010) Neoadjuvant chemotherapy or primary surgery in stage iiic or iv ovarian cancer. *New England Journal of Medicine* **363**,
13 943–953. PMID: 20818904.

[80] Dilruba, S & Kalayda, G. V. (2016) Platinum-based drugs: past, present and future. *Cancer*
14 *Chemother Pharmacol* **77**, 1103–1124.

[81] Dasari, S & Tchounwou, P. B. (2014) Cisplatin in cancer therapy: molecular mechanisms
14 of action. *Eur J Pharmacol* **740**, 364–378.

[82] Johnstone, T. C, Park, G. Y, & Lippard, S. J. (2014) Understanding and improving platinum
14 anticancer drugs–phenanthriplatin. *Anticancer Res* **34**, 471–476.

[83] Institute, N. C. (2014) The "accidental" cure—platinum-based treatment for cancer: The
14 discovery of cisplatin.

[84] Minassian, L. M, Cotechini, T, Huitema, E, & Graham, C. H. (2019) Hypoxia-Induced
16 Resistance to Chemotherapy in Cancer. *Adv Exp Med Biol* **1136**, 123–139.

[85] Tomita, A. (2016) Genetic and Epigenetic Modulation of CD20 Expression in B-Cell Malignancies: Molecular Mechanisms and Significance to Rituximab Resistance. *J Clin Exp Hematop* **56**, 89–99. 16

[86] Damia, G & Broggini, M. (2019) Platinum Resistance in Ovarian Cancer: Role of DNA Repair. *Cancers (Basel)* **11**. 16

[87] Spencer, D. M, Bilardi, R. A, Koch, T. H, Post, G. C, Nafie, J. W, Kimura, K.-I, Cutts, S. M, & Phillips, D. R. (2008) Dna repair in response to anthracycline–dna adducts: A role for both homologous recombination and nucleotide excision repair. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **638**, 110–121. 16

[88] Buonaguro, F. M, Caposio, P, Tornesello, M. L, De Re, V, & Franco, R. (2019) Cancer Diagnostic and Predictive Biomarkers 2018. *Biomed Res Int* **2019**, 3879015. 16

[89] Coppedè, F, Lopomo, A, Spisni, R, & Migliore, L. (2014) Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World J Gastroenterol* **20**, 943–956. 16

[90] van IJzendoorn, D. G. P, Szuhai, K, Briaire-de Bruijn, I. H, Kostine, M, Kuijjer, M. L, & Bovée, J. V. M. G. (2019) Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput Biol* **15**, e1006826. 16

[91] Li, D & Chan, D. W. (2014) Proteomic cancer biomarkers from discovery to approval: it's worth the effort. *Expert Rev Proteomics* **11**, 135–136. 16

[92] Dregely, I, Prezzi, D, Kelly-Morland, C, Roccia, E, Neji, R, & Goh, V. (2018) Imaging biomarkers in oncology: Basics and application to MRI. *J Magn Reson Imaging* **48**, 13–26. 16

[93] McCabe, N, Turner, N. C, Lord, C. J, Kluzek, K, Bialkowska, A, Swift, S, Giavara, S, O'Connor, M. J, Tutt, A. N, Zdzienicka, M. Z, Smith, G. C, & Ashworth, A. (2006) Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Res* **66**, 8109–8115. 16

[94] Ganesan, A, Arimondo, P. B, Rots, M. G, Jeronimo, C, & Berdasco, M. (2019) The timeline of epigenetic drug discovery: from reality to dreams. *Clin Epigenetics* **11**, 174. 17

[95] Derissen, E. J, Beijnen, J. H, & Schellens, J. H. (2013) Concise drug review: azacitidine and decitabine. *Oncologist* **18**, 619–624. 17

[96] Lyko, F & Brown, R. (2005) DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. *J Natl Cancer Inst* **97**, 1498–1506. 17

[97] Amato, R. J. (2007) Inhibition of DNA methylation by antisense oligonucleotide MG98 as cancer therapy. *Clin Genitourin Cancer* **5**, 422–426. 17

[98] Tollefsbol, T. O, ed. (2018) in *Epigenetics in Human Disease (Second Edition)*, Translational Epigenetics. (Academic Press) Vol. 6, Second edition edition, p. ii. 17

[99] Sarkozy, C, Morschhauser, F, Dubois, S, Molina, T, Michot, J. M, Cullières-Dartigues, P, Suttle, B, Karlin, L, Le Gouill, S, Picquenot, J. M, Dubois, R, Tilly, H, Herbaux, C, Jardin, F, Salles, G, & Ribrag, V. (2020) A LYSA Phase Ib Study of Tazemetostat (EPZ-6438) plus R-CHOP in Patients with Newly Diagnosed Diffuse Large B-Cell Lymphoma (DLBCL) with Poor Prognosis Features. *Clin Cancer Res* **26**, 3145–3153.

[100] Bhatla, T, Wang, J, Morrison, D. J, Raetz, E. A, Burke, M. J, Brown, P, & Carroll, W. L. (2012) Epigenetic reprogramming reverses the relapse-specific gene expression signature and restores chemosensitivity in childhood B-lymphoblastic leukemia. *Blood* **119**, 5201–5210.

[101] Cronk, D. (2013) in *Drug Discovery and Development (Second Edition)*, eds. Hill, R & Rang, H. (Churchill Livingstone), Second edition edition, pp. 95 – 117.

[102] Zhang, X. D. (2011) Illustration of ssmd, z score, ssmd*, z* score, and t statistic for hit selection in rnai high-throughput screens. *Journal of Biomolecular Screening* **16**, 775–785. PMID: 21515799.

[103] Ritz, C, Baty, F, Streibig, J. C, & Gerhard, D. (2015) Dose-response analysis using r. *PLOS ONE* **10**.

[104] Yadav, B, Pemovska, T, Szwajda, A, Kulesskiy, E, Kontro, M, Karjalainen, R, Majumder, M. M, Malani, D, Murumägi, A, Knowles, J, Porkka, K, Heckman, C, Kallioniemi, O, Wennerberg, K, & Aittokallio, T. (2014) Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci Rep* **4**, 5193.

[105] Potdar, S, Ianevski, A, Mpindi, J. P, Bychkov, D, Fiere, C, Ianevski, P, Yadav, B, Wennerberg, K, Aittokallio, T, Kallioniemi, O, Saarela, J, & Östling, P. (2020) Breeze: an integrated quality control and data analysis application for high-throughput drug screening. *Bioinformatics* **36**, 3602–3604.

[106] Sanger, F, Nicklen, S, Coulson, A. R, Sanger, F, Nicklen, S, & Coulsen, A. R. (1992) DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology* **24**, 104–108.

[107] Brewster, J. L, Beason, K. B, Eckdahl, T. T, & Evans, I. M. (2004) The microarray revolution: Perspectives from educators. *Biochem Mol Biol Educ* **32**, 217–227.

[108] Hoheisel, J. D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**, 200–210.

[109] Goodwin, S, McPherson, J. D, & McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351.

[110] Stark, R, Grzelak, M, & Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631–656.

[111] van Dijk, E. L, Jaszczyszyn, Y, Naquin, D, & Thermes, C. (2018) The Third Revolution in Sequencing Technology. *Trends Genet* **34**, 666–681.

[112] Wang, Y, Yang, Q, & Wang, Z. (2014) The evolution of nanopore sequencing. *Front Genet* **5**, 449.

[113] Hwang, B, Lee, J. H, & Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 1–14. 22

[114] Frommer, M, McDonald, L. E, Millar, D. S, Collis, C. M, Watt, F, Grigg, G. W, Molloy, P. L, & Paul, C. L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827–1831. 23

[115] Meissner, A, Gnirke, A, Bell, G. W, Ramsahoye, B, Lander, E. S, & Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868–5877. 23

[116] (2012) *Agilent SureSelect Human Methyl-Seq for the Quantitative Analysis of DNA Methylation with Single-Base Resolution*. Agilent Technical Overview, Publication Number 5991-0166EN. 23

[117] Smallwood, S. A, Lee, H. J, Angermueller, C, Krueger, F, Saadeh, H, Peat, J, Andrews, S. R, Stegle, O, Reik, W, & Kelsey, G. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817–820. 23

[118] Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`). 24

[119] Bolger, A. M, Lohse, M, & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. 25

[120] Krueger, F. (2010) Trim Galore! (`http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`). 25

[121] Dobin, A, Davis, C. A, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M, & Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. 25

[122] Krueger, F & Andrews, S. R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572. 25

[123] Xi, Y & Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232. 25

[124] Roberts, A & Pachter, L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* **10**, 71–73. 25

[125] Bray, N. L, Pimentel, H, Melsted, P, & Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527. 25

[126] Barturen, G, Rueda, A, Oliver, J. L, & Hackenberg, M. (2013) MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Res* **2**, 217. 25

[127] Love, M. I, Huber, W, & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550. 25

[128] Assenov, Y, Müller, F, Lutsik, P, Walter, J, Lengauer, T, & Bock, C. (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* **11**, 1138–1140.

[129] Müller, F, Scherer, M, Assenov, Y, Lutsik, P, Walter, J, Lengauer, T, & Bock, C. (2019) RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol* **20**, 55.

[130] Sun, D, Xi, Y, Rodriguez, B, Park, H. J, Tong, P, Meong, M, Goodell, M. A, & Li, W. (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* **15**, R38.

[131] Akalin, A, Kormaksson, M, Li, S, Garrett-Bakelman, F. E, Figueroa, M. E, Melnick, A, & Mason, C. E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**, R87.

[132] Zheng, X, Zhang, N, Wu, H. J, & Wu, H. (2017) Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol* **18**, 17.

[133] Feng, H & Wu, H. (2019) Differential methylation analysis for bisulfite sequencing using DSS. *Quant Biol* **7**, 327–334.

[134] Chen, E. Y, Tan, C. M, Kou, Y, Duan, Q, Wang, Z, Meirelles, G. V, Clark, N. R, & Ma'ayan, A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128.

[135] Kuleshov, M. V, Jones, M. R, Rouillard, A. D, Fernandez, N. F, Duan, Q, Wang, Z, Koplev, S, Jenkins, S. L, Jagodnik, K. M, Lachmann, A, McDermott, M. G, Monteiro, C. D, Gundersen, G. W, & Ma'ayan, A. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–97.

[136] Kamburov, A, Pentchev, K, Galicka, H, Wierling, C, Lehrach, H, & Herwig, R. (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* **39**, D712–717.

[137] Kanehisa, M, Furumichi, M, Tanabe, M, Sato, Y, & Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361.

[138] Jassal, B, Matthews, L, Viteri, G, Gong, C, Lorente, P, Fabregat, A, Sidiropoulos, K, Cook, J, Gillespie, M, Haw, R, Loney, F, May, B, Milacic, M, Rothfels, K, Sevilla, C, Shamovsky, V, Shorser, S, Varusai, T, Weiser, J, Wu, G, Stein, L, Hermjakob, H, & D'Eustachio, P. (2020) The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498–D503.

[139] Slenter, D. N, Kutmon, M, Hanspers, K, Riutta, A, Windsor, J, Nunes, N, M?lius, J, Cirillo, E, Coort, S. L, Digles, D, Ehrhart, F, Giesbertz, P, Kalafati, M, Martens, M, Miller, R, Nishida, K, Rieswijk, L, Waagmeester, A, Eijssen, L. M. T, Evelo, C. T, Pico, A. R, & Willighagen, E. L. (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46**, D661–D667.

[140] Carbon, S, Douglass, E, Good, B. M, Unni, D. R, Harris, N. L, Mungall, C. J, Basu, S, Chisholm, R. L, Dodson, R. J, Hartline, E, Fey, P, Thomas, P. D, Albou, L. P, Ebert, D, Kesling, M. J, Mi, H, Muruganujan, A, Huang, X, Mushayahama, T, LaBonte, S. A, Siegele, D. A, Antonazzo, G, et al. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334. 26

[141] Clough, E & Barrett, T. (2016) The Gene Expression Omnibus Database. *Methods Mol Biol* **1418**, 93–110. 26

[142] Leinonen, R, Sugawara, H, & Shumway, M. (2011) The sequence read archive. *Nucleic Acids Res* **39**, 19–21. 26

[143] Stunnenberg, H. G, Hirst, M, Abrignani, S, Adams, D, de Almeida, M, Altucci, L, Amin, V, Amit, I, Antonarakis, S. E, Aparicio, S, Arima, T, Arrigoni, L, Arts, R, Asnafi, V, Esteller, M, Bae, J. B, Bassler, K, Beck, S, Berkman, B, Bernstein, B. E, Bilenky, M, Bird, A, Bock, C, Boehm, B, Bourque, G, Breeze, C. E, Brors, B, Bujold, D, et al. (2016) The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145–1149. 26

[144] Barretina, J, Caponigro, G, Stransky, N, Venkatesan, K, Margolin, A. A, Kim, S, Wilson, C. J, Lehár, J, Kryukov, G. V, Sonkin, D, Reddy, A, Liu, M, Murray, L, Berger, M. F, Monahan, J. E, Morais, P, Meltzer, J, Korejwa, A, Jané-Valbuena, J, Mapa, F. A, Thibault, J, Bric-Furlong, E, Raman, P, Shipway, A, Engels, I. H, Cheng, J, Yu, G. K, Yu, J, Aspesi, P, de Silva, M, Jagtap, K, Jones, M. D, Wang, L, Hatton, C, Palescandolo, E, Gupta, S, Mahan, S, Sougnez, C, Onofrio, R. C, Liefeld, T, MacConaill, L, Winckler, W, Reich, M, Li, N, Mesirov, J. P, Gabriel, S. B, Getz, G, Ardlie, K, Chan, V, Myer, V. E, Weber, B. L, Porter, J, Warmuth, M, Finan, P, Harris, J. L, Meyerson, M, Golub, T. R, Morrissey, M. P, Sellers, W. R, Schlegel, R, & Garraway, L. A. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607. 26

[145] Lappalainen, I, Almeida-King, J, Kumanduri, V, Senf, A, Spalding, J. D, Ur-Rehman, S, Saunders, G, Kandasamy, J, Caccamo, M, Leinonen, R, Vaughan, B, Laurent, T, Rowland, F, Marin-Garcia, P, Barker, J, Jokinen, P, Torres, A. C, de Argila, J. R, Llobet, O. M, Medina, I, Puy, M. S, Alberich, M, de la Torre, S, Navarro, A, Paschall, J, & Flicek, P. (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* **47**, 692–695. 26

[146] Tomczak, K, Czerwińska, P, & Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**, 68–77. 26

[147] ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640. 26

[148] ENCODE Project Consortium, Moore, J. E, Purcaro, M. J, Pratt, H. E, Epstein, C. B, Shoresh, N, Adrian, J, Kawli, T, Davis, C. A, Dobin, A, Kaul, R, Halow, J, Van Nostrand, E. L, Freese, P, Gorkin, D. U, Shen, Y, He, Y, Mackiewicz, M, Pauli-Behn, F, Williams, B. A, Mortazavi, A, Keller, C. A, Zhang, X. O, Elhajjajy, S. I, Huey, J, Dickel, D. E, Snetkova, V, Wei, X, Wang, X, Rivera-Mulia, J. C, Rozowsky, J, Zhang, J, Chhetri, S. B, Zhang, J, Victorsen, A, White, K. P, Visel, A, Yeo, G. W, Burge, C. B, Lécuyer, E, Gilbert, D. M, Dekker, J, Rinn, J, Mendenhall, E. M, Ecker, J. R, Kellis, M, Klein, R. J, Noble, W. S, Kundaje, A, Guigó, R, Farnham, P. J, Cherry, J. M, Myers, R. M, Ren, B, Graveley, B. R, Gerstein, M. B, Pennacchio, L. A, Snyder, M. P, Bernstein, B. E, Wold, B, Hardison, R. C, Gingeras, T. R, Stamatoyannopoulos, J. A, & Weng, Z. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710. 26

[149] Afgan, E, Baker, D, Batut, B, van den Beek, M, Bouvier, D, Cech, M, Chilton, J, Clements, D, Coraor, N, Grüning, B. A, Guerler, A, Hillman-Jackson, J, Hiltemann, S, Jalili, V, Rasche, H, Soranzo, N, Goecks, J, Taylor, J, Nekrutenko, A, & Blankenberg, D. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**, W537–W544. 27

[150] Wolstencroft, K, Haines, R, Fellows, D, Williams, A, Withers, D, Owen, S, Soiland-Reyes, S, Dunlop, I, Nenadic, A, Fisher, P, Bhagat, J, Belhajjame, K, Bacall, F, Hardisty, A, Nieva de la Hidalga, A, Balcazar Vargas, M. P, Sufi, S, & Goble, C. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41**, W557–561. 27

[151] Kallio, M. A, Tuimala, J. T, Hupponen, T, Klemelä, P, Gentile, M, Scheinin, I, Koski, M, Käki, J, & Korpelainen, E. I. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507. 27

[152] Reich, M, Liefeld, T, Gould, J, Lerner, J, Tamayo, P, & Mesirov, J. P. (2006) GenePattern 2.0. *Nat Genet* **38**, 500–501. 27

[153] Ovaska, K, Laakso, M, Haapa-Paananen, S, Louhimo, R, Chen, P, Aittomäki, V, Valo, E, Núñez-Fontarnau, J, Rantanen, V, Karinen, S, Nousiainen, K, Lahesmaa-Korpinen, A. M, Miettinen, M, Saarinen, L, Kohonen, P, Wu, J, Westermarck, J, & Hautaniemi, S. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med* **2**, 65. 27, 32

[154] Sadedin, S. P, Pope, B, & Oshlack, A. (2012) Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**, 1525–1526. 27

[155] Di Tommaso, P, Chatzou, M, Floden, E. W, Barja, P. P, Palumbo, E, & Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319. 27

[156] Köster, J & Rahmann, S. (2018) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600. 27

[157] Cervera, A, Rantanen, V, Ovaska, K, Laakso, M, Nuñez-Fontarnau, J, Alkodsi, A, Casado, J, Facciotto, C, Häkkinen, A, Louhimo, R, Karinen, S, Zhang, K, Lavikka, K, Lyly, L, Pal Singh, M, & Hautaniemi, S. (2019) Anduril 2: upgraded large-scale data integration framework. *Bioinformatics* **35**, 3815–3817. 32

[158] Chang, W, Cheng, J, Allaire, J, Xie, Y, & McPherson, J. (2020) *shiny: Web Application Framework for R*. R package version 1.5.0. 33

[159] Sievert, C. (2020) *Interactive Web-Based Data Visualization with R, plotly, and shiny*. (Chapman and Hall/CRC). 33

[160] Louhimo, R & Hautaniemi, S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* **27**, 887–888. 34

[161] Therneau, T. M. (2020) *A Package for Survival Analysis in R*. R package version 3.2-7. 34

[162] Terry M. Therneau & Patricia M. Grambsch. (2000) *Modeling Survival Data: Extending the Cox Model*. (Springer, New York). 34

[163] Kassambara, A, Kosinski, M, & Biecek, P. (2020) *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.7. 34

[164] Bisserier, M & Wajapeyee, N. (2018) Mechanisms of resistance to EZH2 inhibitors in diffuse large B-cell lymphomas. *Blood* **131**, 2125–2137. 36

[165] Lefor, A. T. (2011) Computational oncology. *Jpn J Clin Oncol* **41**, 937–947. 41

[166] Groselj, B, Sharma, N. L, Hamdy, F. C, Kerr, M, & Kiltie, A. E. (2013) Histone deacetylase inhibitors as radiosensitisers: effects on DNA damage signalling and repair. *Br J Cancer* **108**, 748–754. 41

[167] Kachhap, S. K, Rosmus, N, Collis, S. J, Kortenhorst, M. S, Wissing, M. D, Hedayati, M, Shabbeer, S, Mendonca, J, Deangelis, J, Marchionni, L, Lin, J, Höti, N, Nortier, J. W, DeWeese, T. L, Hammers, H, & Carducci, M. A. (2010) Downregulation of homologous recombination DNA repair genes by HDAC inhibition in prostate cancer is mediated through the E2F1 transcription factor. *PLoS One* **5**, e11208. 41

[168] Lu, Y, Chan, Y. T, Tan, H. Y, Li, S, Wang, N, & Feng, Y. (2020) Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy. *Mol Cancer* **19**, 79. 41

[169] Roberti, A, Valdes, A. F, Torrecillas, R, Fraga, M. F, & Fernandez, A. F. (2019) Epigenetics in cancer therapy and nanomedicine. *Clin Epigenetics* **11**, 81. 41

[170] Kim, H. J, Maiti, P, & Barrientos, A. (2017) Mitochondrial ribosomes in cancer. *Semin Cancer Biol* **47**, 67–81. 42

[171] Madhavan, S, Beckman, R. A, McCoy, M. D, Pishvaian, M. J, Brody, J. R, & Macklin, P. (2021) Envisioning the future of precision oncology trials. *Nature Cancer* **2**, 9–11. 43